Propensity score adjustment of a treatment effect with missing data in psychiatric health services research

Benjamin Mayer[1], Bernd Puschner[2]

ABSTRACT

BACKGROUND: Missing values are a common problem for data analyses in observational studies, which are frequently applied in health services research. This paper examines the usefulness of different approaches in tackling the problem of incomplete observational data, focusing on whether the multiple imputation (MI) strategy yields adequate estimates when applied to a complex analysis framework.

METHODS: Based on observational study data originally comparing three forms of psychotherapy, a simulation study with different missing data scenarios was conducted. The considered analysis model comprised a propensity score-adjusted treatment effect estimation. Missing values were evaluated using complete case analysis, different MI approaches, as well as mean and regression imputation.

RESULTS: All point estimators of the applied methods fall within the 95% confidence interval of the treatment effect, derived from the complete simulation data set. Highest deviation was observed for complete case analysis. A distinct superiority of MI methods could not be demonstrated.

CONCLUSIONS: Since there was no clear benefit of one method over another in dealing with missing values, health services researchers faced with incomplete observational data are well-advised to apply different imputation methods and compare the results in order to get an impression of their sensitivity.

Key words: Missing values; imputation; propensity score; health services research

(1) Institute of Epidemiology and Medical Biometry, Ulm University, Germany
(2) Section Process-Outcome Research, Department of Psychiatry II, Ulm University, Germany

CORRESPONDING AUTHOR: *Benjamin Mayer, Institute of Epidemiology and Medical Biometry, Ulm University, Schwabstr. 13 - 89075 Ulm, Germany , Telephone (+49) 731 5026896 , Fax (+49) 731 5026902. Email: benjamin.mayer@uni-ulm.de*

DOI: ***10.2427/10214***
Accepted on October 2, 2014

INTRODUCTION

Non-randomized observational studies are common in health services research. These studies must address the problem of biased effect estimates due to confounding and missing values. Propensity scores have been suggested to tackle confounding [1] in order to ensure comparability of observation groups. A propensity score represents the conditional probability that, given the existing covariable structure, a binary response variable has a

specific manifestation. Calculation of propensity scores for two comparison groups via logistic regression is straightforward, and can later be used in different analysis methods (e.g. propensity score matching, multiple regression analysis) to estimate adjusted and unbiased group effects [2].

Due to the lack of uniform recommendations in handling incomplete data, the presence of missing data aggravates the problem of biased parameter estimates in non-randomized studies. There are numerous approaches to deal with missing data. In complete case analysis (CCA), incomplete cases are simply excluded from the analysis. Although this method has several drawbacks [3], it is frequently applied, especially in health services research [4]. Regarding the method of mean imputation, a missing value is replaced by the mean of all observed values. In regression imputation, missing values are replaced by applying a regression model to the data, in which all completely observed variables are the covariates and the incomplete variable is the outcome. Both approaches fall under the category of single imputation (SI) methods since each missing value is replaced once. In contrast, missing data are imputed multiple times when the multiple imputation (MI) strategy is applied. Again, different approaches are available which result in several completed data sets. These are analysed separately using an analysis plan, which originally was intended (e.g. Cox regression) after the imputation is completed. Finally, the resulting parameter estimates of interest (e.g. hazard ratios) from each model are combined into a single MI estimate according to Rubin's rules [5].

There is consensus that SI methods underestimate variability of the data which may lead to biased parameter estimates [6]. MI methods generate more accurate estimates since they take into account additional variability resulting from the imputation itself [5,7], and have thus been recommended in current guidelines [3]. However, MI estimates are only meaningful if missing data are independent from both observed and unobserved values (missing completely at random(MCAR)), or if the absence of data can be fully explained by the observed values (missing at random (MAR)).

Recent efforts in methodological research using typical model frameworks have shown that MI can be easily applied to validate prognostic models [8] and that, compared to other approaches, it is the method of choice for such models [9]. MI was also found to be superior compared to other imputation methods for a propensity score-adjusted effect estimation using simulated data [10]. Additionally, the Expectation–Maximization algorithm (EM) has been used to estimate propensity scores in the presence of missing data [5,11], but it has not been compared to other methods. Furthermore, no differences were found between various imputation methods in order to calculate propensity scores to be used for matching of patients [12]. Likewise, different MI methods produced similar values for the area under the curve (AUC) as a measure of concordance in logistic regression models [13].

Taken together, there is limited and inconsistent knowledge on how to accurately estimate propensity scores in the presence of missing data in observational studies. Moreover, specific realizations of the MI strategy in complex analysis frameworks have not yet been fully investigated. This paper intended to evaluate the application of established imputation methods in a typical analysis situation in health services research based on observational data. The primary hypothesis was that the applied MI strategies produce less biased results than the alternative approaches. A simulation study with real study data from psychiatric health services research investigated different missing data scenarios to assess the ability of all applied imputation approaches to approximate the original results.

METHODS

Data example

TRANS-OP data

The simulation study is based on a longitudinal data set from mental health services research. The study "Transparency and Outcome Orientation in Outpatient Psychotherapy" (TRANS-OP) investigated the course of outpatient psychotherapy over two years. Study participants (N=787) were recruited between 1998 and 2000 in Germany from insurees of a major private health insurance company ("Deutsche Krankenversicherung" (DKV)) and received three forms of psychotherapy: psychodynamic psychotherapy (PD, N=402, 51.1%), cognitive-

behavioural therapy (CBT, N=249, 31.6%), or psychoanalytic psychotherapy (PA, N=136, 17.3%).

TRANS-OP is a naturalistic prospective observational study optimised for the application of hierarchical linear models, including five measurement points over two years. All participants received initial questionnaires (T1) as well as 1 1/2 and two years thereafter (T4 and T5). Intermediate measurement points T2 and T3 were administered randomly at two out of seven possible points in time (4, 8, 16, 26, 40, 52, and 64 weeks from intake). Non-equidistant intervals were chosen to allow for more frequent assessments in the early treatment phase. This design provides a rather fine-graded time grid for the sample (a total of 10 measurements over two years), while at the same time keeping the burden on the individual patient at an acceptable level.

Patient-rated symptomatic impairment was measured from T1-T5 with the German version [14] of Derogatis' [15] Symptom-Check-List (SCL-90-R). This is a widely used self-report scale comprised of 90 items each on a five-point Likert scale ("not at all" ... "very much"), yielding the Global Severity Index (GSI) indicating mean impairment over all 90 items. See [16,17,18] for details on the TRANS-OP study.

Simulation data set

In order to allow for both a straightforward calculation and interpretation of propensity scores for two treatment groups, subjects receiving psychoanalytic psychotherapy were excluded, resulting in a subsample of N=651 patients in PD and CBT. Following the prediction model in Puschner and Kordy [16] variables retained in the simulation data set were symptomatic impairment (SCL-90-R GSI) and the covariates age, gender, professional status, family status, duration of illness and type of psychotherapy. All cases with incomplete information on these variables were excluded, resulting in a final sample size of N=504 patients for the simulations. The intention was to have "true" estimates which can be compared with those arising after missing data have been simulated and imputed again. Mean age of the participants was 43.9 (± 11.2) years and 55% were female. The majority had an academic degree (59%) while 45% were married (31% single). Mean duration of illness was 15.9 months (± 9.7 months).

Simulation study

Analysis model

The analysis strategy for evaluating the simulations included two models. The first one addressed the estimation of the adjusted treatment effect (PD vs. CBT) on the outcome variable symptomatic impairment. Therefore, a multiple linear regression model

$$(1)\ y = \theta_0 + \theta_1 \cdot x_{treat} + \theta_2 \cdot x_{PS}$$

was applied, where y is the outcome and x_{treat} represents the group status (PD or CBT). The treatment effect θ_1 was the primary endpoint for evaluating the simulation results. The adjustment variable x_{PS} in model (1) arose from a second prediction model indicating a subject's propensity score to belong to either the PD or CBT treatment group, respectively. The propensity score was obtained by a logistic regression model

$$(2)\ ln\ (p/1\text{-}p) = \beta_1 \cdot x_{age} + \beta_2 \cdot x_{gender} + \beta_3 \cdot x_{profession} + \beta_4 \cdot x_{duration} + \beta_5 \cdot x_{family},$$

with p denoting the conditional probability of a subject receiving PD, based on the given covariates.

Simulation scenarios and missing mechanisms

Different missing value scenarios included the missing data mechanisms MCAR and MAR, as well as various proportions of missing values (low=5%, moderate=20%, high=50%). 1000 runs were conducted for each simulation scenario, with the continuous covariable of duration of illness chosen to have missing observations. Missing data according to MCAR was generated by randomly selecting the respective proportion of study participants. To create a MAR mechanism, it was determined that female patients are more likely to have a missing value for duration of illness.

Moreover, different strategies for handling missing data were compared: (i) a complete case analysis, (ii) a mean imputation, (iii)

a regression imputation, and (iv) three approaches of the MI strategy, each with m=5 imputations and the application of the Markov Chain Monte Carlo (MCMC) method to obtain the imputation values. The MI estimators of the primary simulation endpoint θ_1 (treatment effect) were constructed in distinct ways (see Figure 1): (1) imputation of missing data and calculation of propensity scores and θ_1 before combining the 5 single parameter estimates, (2) calculation of propensity scores after imputation of missing values and combination of the single propensity score estimates before estimating the treatment effect θ_1, and (3) applying a mean imputation and the logistic regression model subsequently to combine the respective regression coefficients β of the five models, resulting in estimates of the propensity scores and the treatment effect θ_1.

Evaluation of simulation results

To assess which of the applied imputation approaches worked best, the adjusted treatment effect θ_1 together with its 95% confidence interval (CI) was calculated and compared to the results of distinct scenarios. Since the sample data set had a longitudinal design, the estimation of $\theta_1$was realised by means of a hierarchical linear model for repeated measurements [19]. The AUC value of the logistic regression model was used to examine the prognostic quality of the model. All analyses were conducted in SAS®, Version 9.2 (www.sas.com).

RESULTS

Table 1 shows the propensity score-adjusted treatment effects θ_1 for different missing data scenarios averaged over all 1000 simulation runs. The treatment effect based on the original simulation data set of N=504 patients with complete information in all variables was θ_1= 0.003 (p =0.944), indicating no effect of PD compared to CBT with respect to symptomatic impairment.

In general, differences in the accuracy of estimates between the investigated missing data mechanisms MCAR and MAR were small. The extent of deviation between the original and simulation based estimations for θ_1 and the respective 95% CI increased with a higher proportion of missing values. The largest difference was found for CCA with 50% missing data in case of MAR, where $\theta_1$was overestimated by the factor 18. However, there was a considerable overlap of CIs of the simulation results with the CI of the original θ_1 estimator, indicating that this difference was not statistically different. Compared to the CCA, mean and regression imputation results showed higher estimation accuracy. Moreover, the results of both methods suggested no distinct dependency of estimation accuracy and missing data mechanism, even in cases with higher proportions of missing data. In most instances, confidence intervals widened with increasing proportions of missing values. This was especially true for the CCA, mean imputation and regression imputation. The MI approaches predominantly showed slightly wider or even narrower confidence intervals. Of the applied MI methods, the third approach showed the largest deviations from the original θ_1 estimator. Furthermore, SI methods produced more accurate estimations of the treatment effect than the MI alternatives in case of MCAR.

Furthermore, as seen in Table 2, the AUC value for the original simulation data set was 0.634, indicating that 63.4% of the patients

FIGURE 1

INVESTIGATED MI STRATEGIES

$$\begin{array}{lccccccccc} \textbf{1}^{\textbf{st}}\ \textbf{approach}: & \text{MI} & \rightarrow & \beta & \rightarrow & \text{PS} & \rightarrow & \hat{\theta} & \rightarrow & \overline{\theta} \\ \textbf{2}^{\textbf{nd}}\ \textbf{approach}: & \text{MI} & \rightarrow & \beta & \rightarrow & \text{PS} & \rightarrow & \overline{\text{PS}} & \rightarrow & \hat{\theta} \\ \textbf{3}^{\textbf{rd}}\ \textbf{approach}: & \text{MI} & \rightarrow & \beta & \rightarrow & \overline{\beta} & \rightarrow & \text{PS} & \rightarrow & \hat{\theta} \end{array}$$

could be correctly classified (patient receives either PD or CBT treatment) based on their covariable information. Over all simulation scenarios, the AUC estimates were between 0.614 and 0.664. With increasing proportions of missing data, AUC values increased for CCA, but decreased for the other imputation methods.

DISCUSSION

The simulation study examined the impact of different strategies in handling missing data when estimating a propensity score-adjusted treatment effect from observational data. The implemented scenarios did not prove an advantage of MI approaches over SI methods with respect to unbiased effect estimates. The extent of bias was generally low since all point estimators for θ_1 fall within the original 95% CI of the treatment effect. However, the maximally observed deviation in case of 50% MAR analyzed with CCA was considerable, indicating once again that this approach is not recommended when analysing incomplete data sets.

The previously cited conclusions [8,9,10] of a distinct superiority of MI approaches could not be confirmed in general. When compared to simple SI approaches, the applied MI methods overestimated the treatment effect more seriously for scenarios with higher

TABLE 1

SIMULATION RESULTS FOR THE TREATMENT EFFECT AND AUC VALUE

		MCAR				MAR			
MISSING	METHOD	θ_1	LCL	UCL	AUC	θ_1	LCL	UCL	AUC
ORIGINAL RESULTS*		0.003	-0.090	0.096	0.634	0.003	-0.090	0.096	0.634
5%	CCA	0.003	-0.093	0.099	0.635	0.009	-0.088	0.106	0.637
	MEAN	0.003	-0.089	0.097	0.633	0.004	-0.088	0.097	0.631
	REGRESSION	0.003	-0.089	0.097	0.633	0.004	-0.089	0.097	0.631
	MI1	0.004	-0.088	0.098	0.632	0.005	-0.088	0.098	0.630
	MI2	0.004	-0.089	0.098	0.632	0.005	-0.089	0.098	0.630
	MI3	0.011	-0.081	0.103	0.632	0.011	-0.081	0.104	0.631
20%	CCA	0.003	-0.101	0.108	0.640	0.017	-0.086	0.121	0.641
	MEAN	0.005	-0.087	0.099	0.630	0.006	-0.087	0.099	0.626
	REGRESSION	0.005	-0.087	0.099	0.630	0.005	-0.087	0.099	0.627
	MI1	0.007	-0.085	0.101	0.626	0.007	-0.085	0.101	0.625
	MI2	0.008	-0.086	0.101	0.627	0.008	-0.086	0.101	0.625
	MI3	0.012	-0.080	0.104	0.627	0.012	-0.080	0.104	0.625
50%	CCA	0.003	-0.130	0.137	0.658	0.054	-0.077	0.187	0.664
	MEAN	0.009	-0.083	0.102	0.623	0.011	-0.081	0.104	0.614
	REGRESSION	0.009	-0.083	0.103	0.622	0.010	-0.082	0.103	0.615
	MI1	0.012	-0.079	0.105	0.617	0.013	-0.079	0.106	0.614
	MI2	0.013	-0.080	0.106	0.617	0.014	-0.079	0.107	0.614
	MI3	0.012	-0.080	0.105	0.617	0.012	-0.080	0.105	0.614

Notes: *=based on 504 patients with complete data; θ_1=treatment effect, LCL=lower 95% confidence limit for θ_1, UCL=upper 95% confidence limit for θ_1; Methods: CCA=complete case analysis; Mean=mean imputation; Regression=regression imputation; MI=Multiple Imputation alternatives according to Figure 1; AUC=area under the curve (prognostic ability of the propensity score model, AUC in [0,1], 0.5 is worst)

amounts of missing data. However, the extent of bias in all MI approaches was limited to a factor of 4.5, compared to a factor of 18 using CCA. Moreover, the simulations indicated that MI was able to limit inaccuracy of the treatment effect's estimated confidence interval. The analyses showed that the underlying logistic regression model which calculated the propensity scores had only a limited ability to predict a patient's belonging to the PD or CBT group, respectively, given the individual covariable structure (AUC=0.634). The application of distinct imputation methods just slightly affected this finding and therefore confirmed the conclusions of Faris et al. [13].

The performed analyses for this paper focused on a typical analytical approach in health services research, where observational studies are common and propensity scores are thus frequently used. Handling of missing data in the considered analysis framework was challenging since propensity scores were initially calculated and used to estimate an adjusted treatment effect afterwards. Hence, imputation should have been valid to ensure a minimal risk for error propagation. Little research has been conducted on how to specifically implement the MI strategy in complex analysis models in order to get valid estimates. According to the conducted simulations the third approach (see Figure 1) seemed to be slightly inferior, since in cases of low and moderate proportions of missing data the estimated treatment effect was even more biased than the applied SI methods.

Limitations

This simulation study also has some limitations. Generating the MAR mechanism was based on the assumption that female patients are more likely to have a missing value for the variable duration of illness. However, it was found that female patients were treated on average only 30 days longer than male patients (p=0.48). Therefore, differences between MCAR and MAR were less obvious than expected in the presented simulations. The effect of missing data mechanism on estimation accuracy, as described in Jackson et al. [20], may have been stronger if the variables to create MAR were correlated stronger or more strongly. Moreover, the original treatment effect was small (θ_1=0.003) and obviously not significant. The effect of different a priori selectivities of the prognostic model was also not captured or observed in this simulation study.

CONCLUSION

The simulations did not suggest a general superiority of MI methods in the considered analysis model. However, all point estimators of the applied methods fall within the 95% CI of the original treatment effect, which arose from the complete data set. This reduces the significance of this finding, which is solely based on the comparison of original and simulated point estimators. In contrast, the simulations confirmed once again that the application of CCA is not advisable, especially if the proportion of missing values is high. Further research is especially required to investigate the behaviour of different strategies in implementing a MI in combined analysis models, incorporating an assessment of the effects of the above-mentioned limitations. Since a clear superiority of MI in a combined analysis model could not be shown by means of the current simulations, health services researchers faced with incomplete observational data are well-advised not to depend exclusively on the MI strategy, but rather to apply different imputation methods and compare the results in order to get an impression of their sensitivity.

ACKNOWLEDGEMENTS: *Benjamin Mayer was funded by the young scientists programme of the German network 'Health Services Research Baden-Württemberg' of the Ministry of Science, Research and Arts in collaboration with the Ministry of Employment and Social Order, Family, Women and Senior Citizens, Baden-Württemberg. The study "Transparency and Outcome Orientation in Outpatient Psychotherapy" (TRANS-OP) was funded by the Deutsche Krankenversicherung (DKV), Cologne, Germany. Moreover, we would like to thank Dr. Hans Kordy who provided the TRANS-OP data for the conducted simulations as well as Samantha Harris, George Taurman and Sandra Roberts for proofreading the manuscript. The authors have no competing interests.*

References

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983, 70(1): 41-55.

[2] Guo S, Fraser MW. Propensity score analysis: statistical methods and applications. SAGE Publications, Inc., Thousand Oaks, 2010.

[3] European Medicines Agency. Guideline on missing data in confirmatory clinical trials. European Medicines Agency, London, 2010.

[4] Mayer B. Fehlende Werte in der Versorgungsforschung [Missing values in health services research]. Monitor Versorgungsforschung 2012, 3: 39-42.

[5] Rubin D.B. Multiple imputation for nonresponse in surveys. John Wiley & Sons, New York, 1987.

[6] Little RJA, Rubin DB. Statistical analysis with missing data. Second edition. John Wiley & Sons, New Jersey, 2002.

[7] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological Methods 2002, 7(2): 147-77.

[8] Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. Journal of Clinical Epidemiology 2002, 63: 205-14.

[9] Janssen KJM, Vergouwe Y, Donders ART, Harrell FE, Jr., Chen Q, Grobbee DE, Moons KGM. Dealing with missing predictor values when applying clinical prediction models. Clinical Chemistry 2009, 55: 994-1001.

[10] Qu Y, Lipkovich I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. Statistics in Medicine 2009, 28(9): 1402-14.

[11] D'Agostino RB, Rubin DB. Estimating and using propensity scores with partially missing data. Journal of the American Statistical Association 2000, 95(451): 749-59.

[12] Mattei A. Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing. Statistical Methods and Applications 2009, 18: 257-73.

[13] Faris PD, Ghali WA, Brant R, Norris CM, Galbraith PD, Knudtson ML. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. Journal of Clinical Epidemiology 2002, 55(2): 184-91.

[14] Franke GH. Die Symptom-Checkliste von Derogatis – Deutsche Version [Derogatis' symptom checklist – German version]. Hogrefe Press, Goettingen, 1995.

[15] Derogatis LR. SCL-90-R: Self report symptom inventory. In Collegium Internationale Psychiatriae Scalarum (Ed.), Internationale Skalen für Psychiatrie. Weinheim: Beltz. 1986.

[16] Puschner B, Kordy H. Mit Transparenz und Ergebnisorientierung zur Optimierung der psychotherapeutischen Versorgung: Eine Studie zur Evaluation ambulanter Psychotherapie [With transparency and results-orientation to the optimum of psychotherapeutic care: a study to evaluate outpatient psychotherapy]. Psychotherapie, Psychosomatik, Medizinische Psychologie 2010, 60: 350-7.

[17] Puschner B, Kraft S, Kaechele H, Kordy H. Course of improvement over 2 years in psychoanalytic and psychodynamic outpatient psychotherapy. Psychology and Psychotherapy: Theory, Research and Practice 2007, 80: 51-68.

[18] Gallas C, Kaechele H, Kraft S, Kordy H, Puschner B. Inanspruchnahme, Verlauf und Ergebnis ambulanter Psychotherapie: Befunde der TRANS-OP-Studie und deren Implikationen für die Richtlinienpsychotherapie [Utilization, course and outcome of outpatient psychotherapy: Results of the TRANS-OP study and implication for guideline-oriented psychotherapy]. Psychotherapeut 2008, 53(6): 414-23.

[19] Raudenbush SW, Bryk AS. Hierarchical linear models: applications and data analysis methods. Sage, Newbury Park, CA, 2001.

[20] Jackson D, White IA, Leese M. How much can we learn about missing data?: an exploration of a clinical trial in psychiatry. Journal of the Royal Statistical Society 2010, 173(3): 593-612.

*