# Closet Similar Subset Imputation

Macaulay Okwuokenye* Karl E. Peace†

**Abstract**

Classifying patients based on stated reasons for missing outcome from different intercurrent events induces patients' subsets in data from clinical trials. Often, data imputation disregards these patients' subsets. We discuss a non-parametric data imputation method that reflects reasons stated for missing data and hence patients' subsets. This subset imputation method is based on a similarity measure between baseline covariates of patients' subset with missing data and a random closest subset without missing data. An illustration is provided.

**Key Words:** Missing data; Hotdeck imputation; Mahalanobis distance; Discrete data imputation; Nearest neighbor

## 1. Introduction

In clinical trials conducted to evaluate the efficacy of therapeutic agents, patients discontinue treatment for different reasons. For example, in the DEFINE study [1], a study conducted to evaluate the efficacy of dimethyl fumarate (BG12) for the treatment of relapse remitting multiple sclerosis, the number of patients who discontinued study drug due to relapse and disease progression (lack of efficacy) in the placebo, BG12 twice daily, and BG12 thrice daily were 23/180, 7/176, and 7/184, respectively.

Classifying patients based on reasons for missing outcome from different intercurrent events induces patient subsets in clinical trial data. For instance, patients whose missing outcome stem from withdrawal due to adverse/serious adverse event or switch to rescue therapy may represent different patient subsets. Differential dropout rates in time may reflect differences in patients baseline disease severity, and hence different subsets among the dropouts. Often, data imputation disregards these patients' subsets.

We discuss a non-parametric data imputation method that reflects reasons stated for missing data and hence patients' subsets. This subset imputation method is based on a similarity measure between baseline covariates of patients' subset with missing data and a random closest subset without missing data. An illustration is provided.

## 2. Method

### 2.1 Data Structure

We assume that there are non-missing response data on $NM$ patients and missing response data on $M$ patients. The data of the $i$th patient from $NM$ is $(X_i, Y_i)$, where $X_i$ is a vector of covariates, and $Y_i$ is the vector of responses. The data of the $j$th patient from $M$ is $(X_j, .)$, where $X_j$ is a vector of covariates, and "." is the vector of missing responses. The task is to leverage information on $X_i$ and $X_j$ or the association between covariates and $Y_i$ in implementing data imputation. An example is the imputation of missing gadolinium enhancing (Gd) lesions in multiple sclerosis where number of Gd lesions are subject to excess zeroes following effective treatment. Other settings include, but not limited to, a) the imputation of patients' infection status in anti-microbial trials [[5] Ch 12] where if the

*Brio Dexteri Pharmaceutical Consultants

†Jiang-Ping Hsu College of Public Health

drug is effective excess zeroes result from higher proportion healed; and b) the imputation of dental caries data among children and the imputation of number of epileptic seizures. In the case of Gd lesions, dental caries, and epileptic seizures, $Y_i$ is count data, whereas in the case of healing status, $Y_i$ is binary. Beside been discrete, data arising from these settings may be subject to excess zeroes or over dispersion, which present additional challenges for imputation.

### 2.2 Closest Similar Subset Imputation

Under imputation using a subset with minimum distance, the task is to use the $X_i$ and $X_j$ to identify the subset $S$ from $NM$ such that the distance between between $S$ and $M$ is minimum. Once $S$ is found, one would randomly select the $Y_i$ of $S$ and use those to impute the missing response values of $M$. This would require one to form the distance between the covariates of $M$ and the covariates of each of the combination $NM$ taken $M$ at a time. When the number of all combination of $NM$ taken $M$ at a time is not computationally difficult, one could enumerate all such possible combinations. When this might be computationally difficult, one could randomly choose a subset of the possible combinations that the computer can handle and use (at the heart of population inference based on random sample from the population) that. Alternatively, one could randomly choose a number of combination much smaller than the computer could handle, compute the distance, and repeat many times generating a sampling distribution of the minimum distance. One could then select any of the subset in the lower percentile (10th say).

In particular, suppose the non-missing dataset $NM$ has 10 patients and the missing dataset $M$ has 2 observations. Further, suppose that $X_1$ and $X_2$ are covariates upon which one wants to match a subset of $NM$ patients to those of $M$. Theoretically, one could form $S$ subsets from $NM$ which equals the combination of 10 things taken 2 at a time [$\binom{10}{2}$= 45]. Then one would compute the distance between the 2 patients in $M$ and each of the 45 subsets of $NM$ in terms of $X_1$ and $X_2$. Then choose the subset $S_{MIN}$ from the subsets in $NM$ that has the smallest distance to $M$. One would then randomly assign missing patients in $M$ to the categories of outcome (numbers of lesions in the running example) proportionate to the percentages of outcome (numbers of Gd lesions in the running example) in $S_{MIN}$.

Suppose the baseline covariate vectors on the $i$th and $j$th patients in the $M$ group are $(x_i, y_i)$ and $(x_j, y_j)$; and that the covariate vectors on the $r$th and $s$th patients in the $NM$ groups are $(x_r, y_r)$ and $(x_s, y_s)$ (where there are 45 such pairs in the above example). These two groups can be visualized in the $XY$ plane. For the $M$ group there are two pairs of points. For the $NM$ group there are 45 pairs of points (but one is forming the distance between the pair of points in the $M$ group to each of the 45 pairs of points in the $NM$ group). So the question is what is a measure of distance between the two sets each consisting of two pairs of points? There are several formulations (any point on the circle passing through the two pairs of points with center the midpoint between the pair and diameter equalling the distance between the pair of points in each subset). In the present case, we used a midpoint that is unique.

Denote $p$ dimensional vector sampled from two preferable, but not necessarily, normal population $T_1$ and $T_2$ by $X_i(i = 1, 2, \ldots m)$ and $Y_i(i = 1, 2, \ldots n)$, respectively. Let $\bar{x}$, $\bar{y}$, $\bar{S_x}$, $\bar{S_y}$ represent, respectively, the mean vectors and covariance matrices. Denote $S_p = (S_x + S_y)/(n + m - 2)$. Mahalanobis distance between the two population is:

$$MD_{sub}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = \sqrt{(\bar{\mathbf{x}} - \bar{\mathbf{y}})^T S_p^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}})}. \quad (1)$$

## 3. Illustration

The confidential nature of the data in the present application precluded actual clinical data presentation; therefore, simulated data are used for illustration. To allow comparability, we demonstrate closest similar subset imputation using a simulated data set that mirrors that used in the COUNTIMP R package [2]. In particular, we generated a data set that contains over dispersed dependent count variable following negative binomial and three continuous predictors, $x_1 - x_3$, that follow $N(0, 1)$ with $\beta_0 - \beta_3$ being the corresponding regression coefficients, and $\beta_0$ being the intercept coefficient. The values of the parameters are $\beta_0 = 1$, $\beta_1 = 0.75$, $\beta_2 = -0.25$, and $\beta_3 = 0.50$. The sample size was 1,000 to be consistent with the sample size generally used in clinical trial. Missing at random was introduced by sampling and declaring missing 20% of the data points where the value of the predictors are less than the mean. See [2]. The combination of 1,000 things taken 200 at a time is enormous for the computer to handle; therefore, we generated 50,000 random subsets of such combinations. We then computed the MD between the covariates of the 200 patients having missing outcome with each of the 50,000 subsets; we selected the subset with minimum distance and used the outcome for this subset with minimum distance for imputation as described above.

Table 1 displays the maximum likelihood parameter estimates from full data set, closest similar subset imputation, full conditional specification imputation, and complete case. The parameter estimates from the closest similar subset imputation are generally similar to those from FCS imputation, exception being that the standard error from the FCS is considerably smaller. This might be due larger sample size (5,000 versus 1,000) used to demonstrate FCS imputation in the COUNTIMP R package. We were unable to apply FCS implemented in the COUNTIMP R package possibly due the incompatibility between the recent version of MICE package and COUNTIMP package.

**Table 1**: Maximum Likelihood Parameter Estimates from Full Data, Closest Similar Subset Imputation, and FCS

| | Full Data Set | | Subset Imputation | | FCS Imputation[1] | | Complete Case | |
|---|---|---|---|---|---|---|---|---|
| N | 1,000 | | 1,000 | | 5,000 | | 800 | |
| Parameter | $\hat{\beta}$ | $\hat{se}$ | $\hat{\beta}$ | $\hat{se}$ | $\hat{\beta}$ | $\hat{se}$ | $\hat{\beta}$ | $\hat{se}$ |
| $\beta_0$ | 1.0121 | 0.0307 | 1.0240 | 0.0301 | 1.0142 | 0.0151 | 1.0106 | 0.0338 |
| $\beta_1$ | 0.7613 | 0.0301 | 0.7648 | 0.0293 | 0.7617 | 0.0154 | 0.7820 | 0.0327 |
| $\beta_2$ | -0.2517 | 0.0284 | -0.2743 | 0.0269 | -0.2208 | 0.0166 | -0.2677 | 0.0307 |
| $\beta_3$ | 0.4892 | 0.0299 | 0.5068 | 0.0295 | 0.4723 | 0.0148 | 0.5026 | 0.0333 |

Note: 1 Results from Kleinke and Reinecke [3].
Note: FCS is full conditional specification; Subset Imputation is closest similar subset imputation.

## 4. Concluding Remarks

The choice of imputation method warrants careful thought because it impacts the implied research question. Irrespective of the sophistication of any imputation approach taken to address missing data problem, no single imputation approach can overcome the limitation of not having complete data. Accordingly, efforts should be invested to avoid missing data using study design and data collection procedures [4]. Sensitivity analyses should be a critical part of missing data imputation. Contextual assumption of and interpretation of results from statistical analysis of imputed data should be clearly stated.

## References

[1] Douglas L. Arnold, Ralf Gold, Ludwig Kappos, Amit Bar-Or, Gavin Giovannoni, Krzysztof Selmaj, Minhua Yang, Ray Zhang, Monica Stephan, Sarah L. Sheikh, and Katherine T. Dawson. Effects of delayed-release dimethyl fumarate on mri measures in the phase 3 define study. *Journal of Neurology*, 261:1794–1802, 2014.

[2] Kristian Kleinke and Jost Reinecke. countimp 1.0 – a multiple imputation package for incomplete count data. Technical report, University of Bielefeld, Faculty of Sociology, 2013.

[3] Kristian Kleinke and Jost Reinecke. Multiple imputation of incomplete zero-inflated count data. *Statistica Neerlandica*, 67(3):311–336, 2013.

[4] National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education.* Washington, DC:, 2010.

[5] Karl E. Peace and Ding-Geng (Din) Chen. *Clinical Trial Methodology*. Chapman and Hall/CRC, 2010.