Research Article

Estimation of Health-Related Quality of Life in the Presence of Missing Values in EQ-5D

Muhammad Atif,[1] Qamruz Zaman,[1] Muhammad Shafiq,[2] and Abdel-Salam G. Abdel-Salam [ORCID][3]

[1]*Department of Statistics, University of Peshawar, Peshawar, Pakistan*
[2]*Institute of Numerical Sciences, Kohat University of Science and Technology, Kohat, Pakistan*
[3]*Department of Mathematics, Statistics and Physics, College of Arts and Science, Qatar University, Doha, Qatar*

Correspondence should be addressed to Abdel-Salam G. Abdel-Salam; abdo@qu.edu.qa

One of the notes worthy problems in analysis of clinical and observational studies is missing data and nonresponse from patients. Turning a blind eye to the missing behavior may provide biased results with overestimated standard errors. The potential impact of the problem may even have more severe impression in estimating health-related quality of life index. This index is an important indicator, widely used in clinical trials for assessing effectiveness of available interventions. Amongst many available measures for estimation of the index, the most rising approach is the EQ-5D preference-based health classifier. This study suggests a cluster-based heuristic algorithm for imputation of missing values in the EQ-5D health classifier to overcome the said problem. The use of auxiliary variable and other dimension's values as evidences increases the chance of correct identification of the missing value and hence makes it unbiased. Comparisons of bootstrap samples suggest that it overcomes the problem of standard errors and provides efficient estimates.

1. Introduction

Provision of medical intervention and clinical facilities on an affordable expense to population is one of the prime goals of public health policy and practice. For this purpose, public health officials use cost-effective ratio to measure the consequence of intervention on physical and mental health of individuals, as well as the additional cost to be paid for improved health conditions. Health-related quality of life (HRQol) is one related concept that is used for comparing the effectiveness of available interventions [1–3]. Many schemes are offered for calculation of HRQol, but a standardized and simplest approach is the EQ-5D preference-based health classifier [4–7]. In this system, health status of an individual is attained by the instrument in a number of dimensions, describing physical and mental fitness. These dimensions include mobility, self-care activities, usual activities, pain/discomfort, and anxiety/depression. Each

dimension of the classifier is presented on the questionnaire with three ordinal levels of responses, i.e., no problem, some/moderate problem, and extreme problem [8, 9]. In this way, the EQ-5D self-classifier provides 243 different possible categories of the health profile. In addition to the EQ-5D classifier, the valuation of HRQol comprises an optical scale as well, usually the visual analogue scale (VAS) or time-trade-off (TTO) scale. Valuations of this visual scale are regressed on the EQ-5D health state vector, and HRQol index is estimated from regression coefficients. The index-based score is typically interpreted along a continuum, where 1 represents the best and 0 represents the worst possible health state [10, 11].

Amongst a number of implications that clinical researchers experience is the problem of missing cases. Most often, the patients miss their appointments due to one reason or the other, and researchers lose their follow up. This phenomenon of nonresponse from the patients may not be

overlooked because the missing part may be informative and can lead to some valuable findings. Dropping patients with missing observations may lead to a misrepresentative finding of the study. But so far, no definite technique is pointed out to be worked in case of missingness in clinical trials [12], particularly in estimation of HRQol. Using the dataset with missing observations may even have more adverse effects on the estimation of HRQol index. In case missing data are informative, the resultant HRQol would be biased with overestimated standard errors [13, 14]. Overall, this study aims to study the impact of missing in the EQ-5D health classifier on HRQol index and suggest a technique for imputation that can overcome the problem. Specifically, this study aims to investigate the impact of deleting cases that have missing observations in the EQ-5Dhealth classifier, introduce an alternate imputation technique by clustering the data on some covariates, and compare the results with some well-known imputation techniques.

1.1. Categorization of Missing Data.
Catalogue of missing data often comprises of three types, i.e., missing not at random (MNAR), missing at random (MAR), and missing completely at random (MCAR) [15]. Though the practical meanings of these three important terms are ambivalent, yet they have some statistical definitions. When the probability of an individual being missing is same and missing cases are considered as the random subsample of population under the study, the type is considered as MCAR. Unlike MCAR, the MAR occurs when phenomena of an individual being missing depends on some information that have already been observed. In these both cases, missing data can be ignored, and these observations can be omitted from the dataset. When the missing observation is related to the value of unobserved data, i.e., the probability of being missing depends on the observation itself, and then, it is denoted as MNAR. The MNAR category is called informative missingness, where the lost part contains some information about response. As a result, the obtained sample is biased, and missing observation cannot be ignored [16, 17].

In case data are ranked on the EQ-5D health classifier, missing behavior may be considered MNAR, as patient with higher pain and anxiety will be less likely to report their health status. Similarly, patients with improved health conditions as a result of intervention avoid visit health practitioner for a follow-up study and hence have small chance of being recorded. As a result, the observed sample would be biased, and some informative parts may be ignored.

1.2. Methods of Dealing with Missing Data.
Several techniques are proposed for imputation of missing values and nonresponse in a dataset. The most common approaches are discussed as follows.

1.2.1. Complete Case Analysis.
In past years, complete case analysis (CCA) has been considered as the ultimate traditional way of dealing with datasets containing missing observation on some attributes. According to this approach, any case with missing observation on some variables is omitted from the data left with only complete cases in the analysis [18]. It is most popular technique because of its ease, and most of statistical packages implement it as default options. However, CCA exclude the complete data on a case that has missing values on some variables. Because of this loss of information, the CCA produce biased estimates of the population parameters. To overcome this, pairwise deletion was introduced which use the pair of variables for which data are available [19].

1.2.2. Single Value Imputation.
Conceivably, the simplest approach to deal with the missing value is to replace it with mean of the observed values for the respective variable. This strategy severely underestimates the standard error, as it does not add much information to the datasets but only increases the sample size. Possibly, mean imputation has some serious problems in replacing the missing values, so the researchers try using the linear regression model and predict its value on the basis of other available variables. The already existing variables are used to predict the value of missing case, consider it to be the true value, and impute it in the dataset. In regression imputation, the imputed value is somehow related to the information available on that particular variable, but the problem of standard error remains the same [20].

1.2.3. Hot-Deck Imputation.
Hot-deck imputation (HDI) is one of the widely used techniques in practice, for handling cases with missing values on some attributes. According to this approach, the missing values are replaced by the observed values from donor's pool that have similar characteristics to the recipient on attributes observed for both. The donor pools are created based on auxiliary variables that are observed for both cases, i.e., respondents and nonrespondents. Andridge and Little [21] reviewed the available literature on statistical properties of HDI and its different invariants. According to them, HDI does not assume statistical distribution or the underlying model as other parametric imputations do. Though hot-deck imputation is intuitive, yet it suffers from a number of limitations. Amongst several, the most challenging drawbacks is that in case of multivariate missing data, the donor cases may not be representative of the recipients.

2. Methods and Materials

In this study, an attempt has been made to investigate the effect of missingness in clinical trials, and a novel algorithm is suggested for imputation of missing values. The general layout of this study is as follows.

In Section 3, a novel cluster-based imputation technique is presented, which can be used for handling missing cases in the EQ-5D health classifier. In Section 3.4, analysis of the complete dataset is carried out and HRQol is estimated for participants of the survey. In Section 3.5, some missing values were generated in the dataset using MCAR to examine impact of missingness on HRQol index. Bootstrap samples were generated from the incomplete dataset and

results were compared. In Section 3.6, the missing values were estimated from MI, HDI, and our novel algorithm to compare the performance of each imputation technique.

2.1. Survey Instruments and Data Collection.

A face-to-face interview was conducted at various public sector hospitals of Peshawar, Pakistan, to obtain the responses of patients at the EQ-5D health state classifier and time-trade-off scale. To ensure randomness, data were collected from 325 patients using the systematic random sampling technique. Along with this information, data on covariates such as "Age of disease," "Age of patient," "Gender," and "Area of residence" were collected.

2.2. Related Work.

Rubin and Schenker [22] proposed the idea of multiple imputations (MI) in clinical trials, where more than one value is to impute for each missing case, estimated from an appropriate probability distribution. Statistical analyses are carried out on each of the resulting dataset and are then combined in order to take a final inferential result into account. If Q_i is the estimate of the ith missing value with associated variance of V_i, then the final estimate of Q would be

$$\overline{Q} = \frac{1}{m} \sum_{i=1}^{m} Q_i, \tag{1}$$

and the associated total variance is

$$V = \overline{v} + \left(1 + \frac{1}{m}\right) B, \tag{2}$$

where $\overline{v}$ is within imputation variability, B is between imputation variability, and m is the number of missing values.

Multiple imputations are generated by the linear regression model, which requires the assumption of multivariate normality. So, this technique might not work in case of categorical response variable. As single value imputation, approximate Bayesian bootstrap (ABB) [23] and fractional hot-deck imputation (FHI) methods were suggested [24]. The ABB method first randomly draws r values with replacement from the r observed values $Y1, \ldots, Yr$ to create Y_{obs}^* and then randomly draws m values with replacement from Y_{obs}^* as imputed values for the m missing values in the target variable Y. The ABB method draws imputations from a resample of the observed data instead of drawing directly from the observed data. This extra step introduces additional variation, which makes the ABB method approximately "proper" for multiple imputations according to Rubin's theory [25]. On the other hand, FHI replaces missing values with a set of imputed values having similar characteristics but assigning weights to it. The simulation studies showed that FHI overcomes the problem of standard errors and produces better results [26].

3. Clusters-Based Multiple Imputation Technique

This study suggests a novel algorithm for imputation of missing values in the EQ-5D health classifier, while estimating HRQol index. Information on some auxiliary covariates is utilized to cluster the dataset with missing observations into various donor groups. If there are "r_i" respondents amongst which "m_i" are missing in i^{th} donor class, then "$m_i/r_i * n$" bootstrap samples of size "r_i" are to be drawn from the respective pool. The Naïve Bayes classifier is applied to each bootstrap sample using other dimension values of EQ-5D as evidences, and "$m_i/r_i * n$" values are estimated for each missing case. The mode of all these estimated values in bootstrap samples is considered as imputation and is replaced instead of missing observation. The general procedure of this method is explained in the following.

3.1. Step1.

Usual K means clustering is performed for segmentation of the dataset into various donors' pools using some appropriate observable covariates. This segmentation of data into homogenous donor pools will identify the pattern of missingness, as patients with low HRQol have a higher chance of not responding to certain questions such as pain, anxiety, and discomfort.

3.2. Step 2.

To ensure randomness and remove bias from imputation, bootstrap samples are generated, and multiple values are to be estimated for each missing value. The average value (mode) of all these multiple imputations is filled up in place of missing observation.

3.3. Step 3.

Finally, the Naïve Bayes classifier is applied to each bootstrap sample in order to classify the missing value to one of the five categories. Known values of the same case in other four dimensions are used as prior evidences for calculating the Bayes probabilities. The posterior probability of observation i that belongs to level k is calculated by

$$\tau_{ik} = pr\left(x_i \in \text{level}_k\right) = \frac{pr\left(x_i | x_i \in \text{level}_k\right) * pr\left(x_i \in \text{level}_k\right)}{pr\left(x_i\right)}. \tag{3}$$

For a missing value, τ_{ik} is calculated for each of the five levels and is assigned to group having maximum posterior probability. Then, the mode of bootstrap samples is used as an imputation of the missing value. The use of Naïve Bayes makes this algorithm more robust by utilizing the information obtained on other dimensions of the EQ-5D health classifier. Figure 1 demonstrates the framework of our proposed algorithm.

3.4. Complete Data Analysis.

TTO scale values are regressed on the EQ-5D preference-based health classifier and the valuation tariffs are estimated by fitting the ordinary least square regression model. According to Table 1, the regression model estimated from complete data is given by

$$\begin{aligned} TTO = {} & 0.089m_1 + 0.01m_2 + 0.02s_1 + 0.035s_2 + 0.157l_1 \\ & + 0.109l_2 + 0.031a_1 + 0.028a_2 + 0.116p_1 + 0.045p_2, \end{aligned} \tag{4}$$
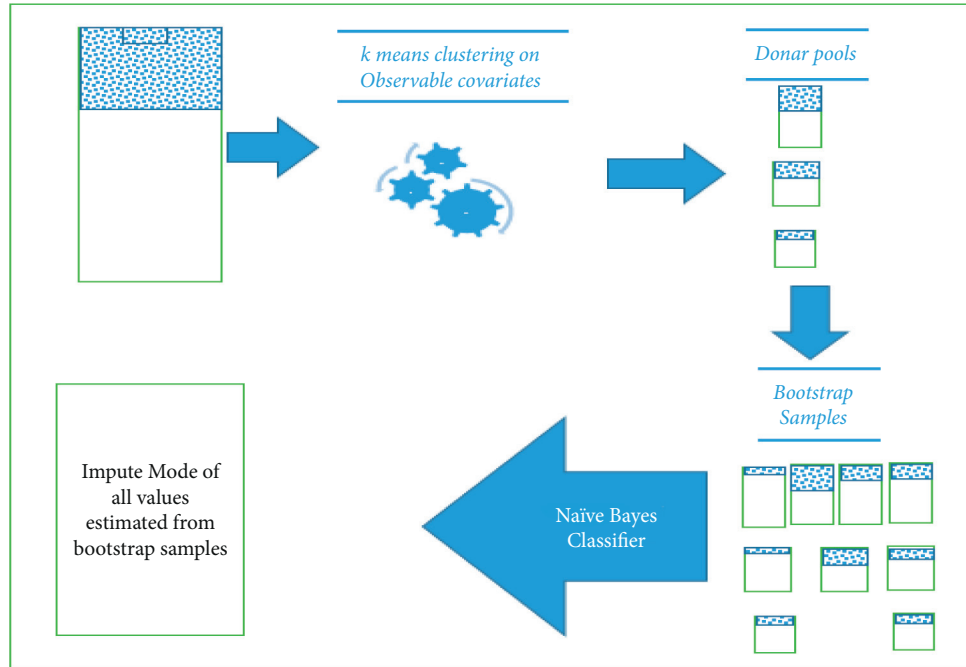
FIGURE 1: Proposed algorithm.

TABLE 1: Valuation tariffs.

Category	Complete data	CCA (±S.E)	MI (±S.E)	HDI (±S.E)	Proposed algorithm (±S.E)
M2	0.089	0.082 (±0.0958)	0.077 (±0.0997)	0.086 (±0.0652)	0.089 (±0.0310)
M3	0.010	0.009 (±0.1308)	0.005 (±0.1203)	0.008 (±0.0953)	0.010 (±0.0602)
S2	0.020	0.014 (±0.0958)	0.013 (±0.0890)	0.016 (±0.0623)	0.020 (±0.0289)
S3	0.035	0.032 (±0.1223)	0.030 (±0.1135)	0.034 (±0.1007)	0.034 (±0.0557)
L2	0.157	0.190 (±0.082)	0.170 (±0.0718)	0.168 (±0.0591)	0.157 (±0.0361)
L3	0.109	0.106 (±0.0997)	0.115 (±0.09281)	0.107 (±0.8203)	0.109 (±0.04121)
P2	0.031	0.010 (±0.0872)	0.022 (±0.0750)	0.030 (±0.0670)	0.031 (±0.04632)
P3	0.028	0.021 (±0.1066)	0.021 (±0.0973)	0.024 (±0.0881)	0.028 (±0.0591)
A2	0.116	0.110 (±0.104)	0.108 (±0.0958)	0.118 (±0.0892)	0.117 (±0.0538)
A3	0.045	0.035 (±0.05729)	0.032 (±0.0511)	0.040 (±0.0448)	0.045 (±0.02931)

where m, s, l, a, and p denote mobility, self-care, usual, anxiety, and pain dimensions, respectively. The subscripts 1 and 2 represent "some/moderate problem" and "extreme problem" in respective dimensions. Valuation tariffs are subtracting from the full health value of $\beta_0 = 1$ in order to estimate HRQol for all patients. Similar regression models are fitted for CCA, MI, and HDI and imputation through our proposed algorithm (cluster-based multiple imputation). The average HRQol for patients is 0.7300 with standard deviation 0.069.

3.5. *Complete Case Analysis.* After fabricating missing values, by deleting 30% responses, the regression model is fitted to only compete cases in each bootstrap sample generated from resultant data. Figure 2 clearly illustrates that most of the times, valuation tariffs (coefficients of the regression model) are underestimated with very large dispersion amongst them.

This amount of bias introduced in valuation tariffs because of missing values led a fake rise in the of HRQol index as presented in Figure 3. The HRQol index is largely over estimated (mean = 0.8037; SD = 0.1407) as a result of CCA applied to the bootstrap samples generated from data with missing cases.

3.6. *Imputation of Missing Values.* Ultimately, the missing values produced in the dataset were imputed using MI, HDI, and our proposed algorithm. Donor pools were formed by clustering the dataset on covariate "age of disease." As suggested in Table 1, though MI reduces standard error of valuation tariffs by a small amount, yet it increases the bias in it. This is due to the fact that missing values in clinical trials are always informative, as those patients who recover their health is less likely to visit the doctor, while those with worst health conditions prefer to change the medicines. Mode imputation ignores this information and replaces the missing value by the average of data, which only increase the sample size but do not add any additional information. HDI slightly improves the results by imputing missing values from similar patients but are still biased. For that reason, our
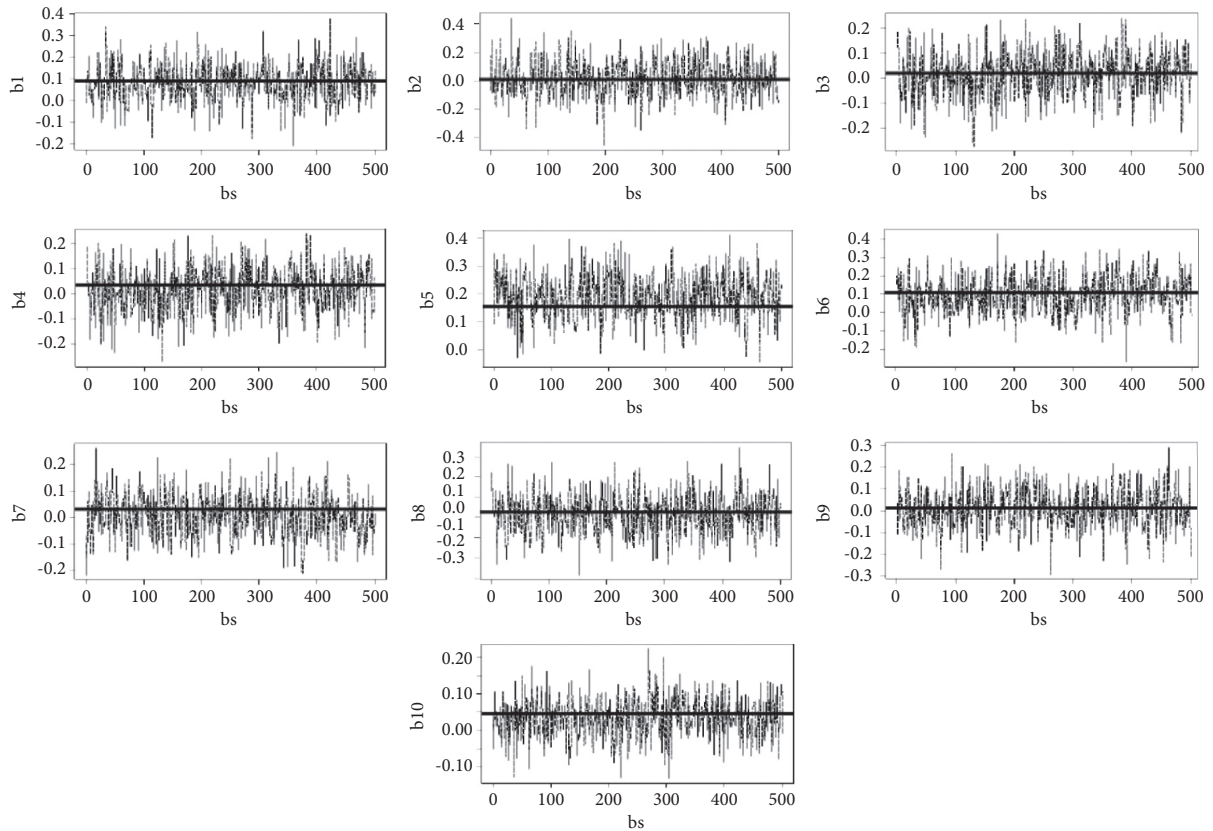
FIGURE 2: Coefficients of the regression model from CCA and complete data. Bold line denotes regression coefficients from complete data; dashed lines represent coefficients estimated through CCA from bootstrap samples after fabricating missing values.
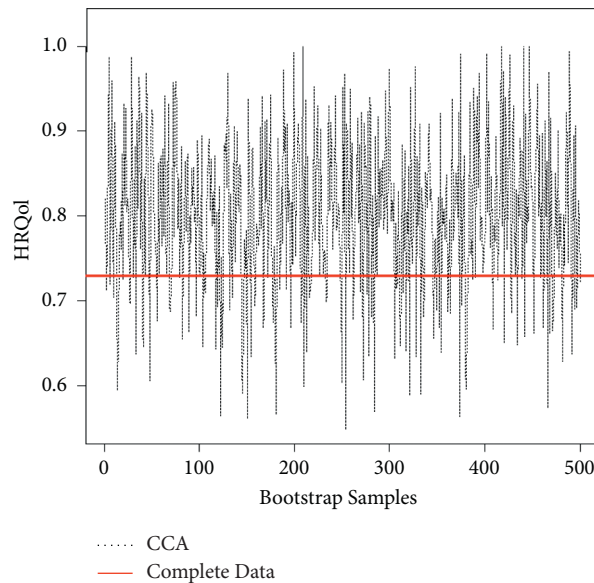

.... CCA
——— Complete Data

FIGURE 3: HRQol from CCA and complete data.

proposed algorithm clusters the dataset by utilizing information obtained from pertinent covariates, and at the same time, other dimensions of the EQ-5D health classifier is used as prior (evidences) in Naïve Bayes posterior probability calculations. These two additional steps succors in identifying the correct health status of patients and not only reduce the variation but also remove the bias introduced as a result of missingness.

Figure 4 illustrates that MI provides highly overestimated HRQol index with large dispersion amongst them.
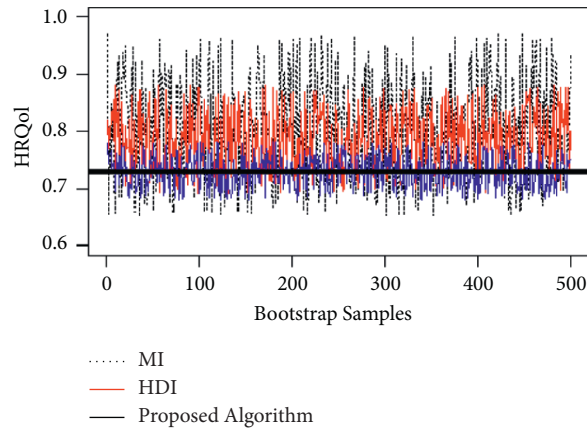
FIGURE 4: HRQol from MI, HDI, and proposed algorithm.

This fake rise in HRQol index is the result of bias involved in estimation of valuation tariffs by replacing the average value instead of the missing value. HDI minimizes the bias to some extend, but still it is not a well representative of actual HRQol index. On contrary, more stable results of HRQol index over repeated samples are obtained when the missing values are imputed by our novel algorithm. The HRQol index raised to 0.8482 and 0.7865, when estimated from MI and HDI, respectively, as compared to 0.7300 from the complete dataset, while from our algorithm, it is 0.7303.

4. Conclusion

Missing data in clinical studies is a common practice and no definitive techniques work best in its presence. In this study, a cluster-based multiple imputation technique is proposed for filling missing values in the EQ-5D preference-based health classifier used for estimation of health-related quality of life. This algorithm tries to estimate multiple values for the missing value using some observable covariates. More advocate and reliable results were obtained in cluster-based multiple imputation than complete case analysis, single value imputation, and hot-deck imputation, when used for estimating missing values in the EQ-5D preference-based health classifier.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

This study was presented in a conference as "Conference: Joint Conference on Biometrics and Biopharmaceutical Statistics" according to the following link. https://www.researchgate.net/publication/319358559_Estimation_of_Health_Related_Quality_of_Life_in_Presence_of_Missing_values_in_EQ-5D.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] M. B. Harrison, G. B. Browne, J. Roberts, P. Tugwell, A. Gafni, and I. D. Graham, "Quality of life of individuals with heart failure," *Medical Care*, vol. 40, no. 4, pp. 271–282, 2002.

[2] W. M. Hopman, M. B. Harrison, H. Coo, E. Friedberg, M. Buchanan, and E. G. VanDenkerkhof, "Associations between chronic disease, age and physical and mental health status," *Chronic Diseases in Canada*, vol. 29, no. 3, pp. 108–117, 2009.

[3] N. Gusi, P. R. Olivares, and R. Rajendram, "The EQ-5D health-related quality of life questionnaire," in *Handbook of Disease Burdens and Quality of Life Measures*, V. R. Preedy and R. R. Watson, Eds., Springer, New York, NY, USA, pp. 87–99, 2010.

[4] M. B. Harrison, I. D. Graham, J. Logan, C. Toman, and E. Friederg, "Evidence to practice: pre-post-implementation study of a patient/provider resource for self-management with heart failure," *International Journal of Evidence-Based Healthcare*, vol. 5, no. 1, pp. 92–101, 2007.

[5] F. Zhang, S. Y. Fu, M. Wang, and J. Du, "Correlation empirical research on the quality of life of Chinese patients with acute ischemic stroke measured by EQ-5D," *China Journal of Pharmaceutical Economics*, vol. 6, pp. 5–12, 2009.

[6] E. Horowitz, I. Abadi-Korek, M. Shani, and J. Shemer, "EQ-5D as a generic measure of health-related quality of life in Israel: reliability, validity and responsiveness," *The Israel Medical Association Journal : IMAJ*, vol. 12, no. 12, pp. 715–720, 2010.

[7] D. A. Revicki and W. R. Lenderking, "Methods and issues associated with the use of quality-adjusted life-years," *Expert Review of Pharmacoeconomics & Outcomes Research*, vol. 12, no. 1, pp. 105–114, 2012.

[8] R. Rabin and F. Charro, "EQ-5D: a measure of health status from the EuroQol Group," *Annals of Medicine*, vol. 33, pp. 337–343, 2001.

[9] K. Burstorm, M. Johannesson, and F. Diderichsen, "Swedish population health related quality of life results using the EQ-5," *Quality of Life Research*, vol. 10, pp. 621–625, 2001.

[10] A. Tsuchiya, S. Ikeda, N. Ikegami et al., "Estimating an EQ-5D population value set: the case of Japan," *Health Economics*, vol. 11, pp. 341–353, 2002.

[11] P. Dolan, "Modeling valuations for EuroQol health states," *Medical Care*, vol. 35, no. 11, pp. 1095–1108, 1997.

[12] D. D. James, A. P. Lori, Z. Qing, F. Zhixuan, and P. Peter, "Analysis of categorical data," *Design and Analysis of Clinical Trials*, vol. 86, pp. 343–388, 2013.

[13] J. L. Roderick, D. Ralph, L. C. Michael et al., "The prevention and treatment of missing data in clinical trials," *New England Journal of Medicine*, vol. 367, no. 14, pp. 1355–1360, 2012.

[14] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 72, pp. 359–364, 1976.

[15] J. W. Graham, "Missing data analysis: making it work in the real world," *Annual Review of Psychology*, vol. 60, no. 1, pp. 549–576, 2009.

[16] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: a gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[17] B. Alma, Pedersen, M. M. Ellen et al., "Missing data and multiple imputation in clinical epidemiological research," *Clinical Epidemiology*, vol. 9, pp. 157–166, 2017.

[18] J. L. Roderick, "Regression with missing X's: a review," *Journal of the American Statistical Association*, vol. 87, no. 420, pp. 1227–1237, 1992.

[19] C. K. Enders, "The performance of the full information maximum likelihood estimator in multiple regression models with missing data," *Educational and Psychological Measurement*, vol. 61, no. 5, pp. 713–740, 2001.

[20] S. V. Buuren, "Flexible imputation of missing data," *Interdisciplinary Statistics Series*, Chapman & Hall/CRC, Florida, FL, USA, 2012.

[21] R. R. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response," *International Statistical Review*, vol. 78, no. 1, pp. 40–64, 2010.

[22] D. B. Rubin and N. Schenker, "Multiple imputations in health care databases: an overview and some applications," *Statistics in Medicine*, vol. 91, pp. 473–489, 1991.

[23] D. B. Rubin and N. Schenker, "Multiple imputations for interval estimation from simple random samples with ignorable non-response," *Journal of the American Statistical Association*, vol. 81, pp. 366–374, 1986.

[24] J. K. Kim and W. Fuller, "Fractional hot deck imputation," *Biometrika*, vol. 91, no. 3, pp. 559–578, 2004.

[25] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, NY, USA, 1987.

[26] J. K. Kim and W. Fuller, "Fractional hot deck imputation," *Biometrika*, vol. 91, no. 3, pp. 559–578, 2004.