PharmaSUG 2013 - Paper SP03

Combining Analysis Results from Multiply Imputed Categorical Data

Bohdana Ratitch, Quintiles, Montreal, Quebec, Canada
Ilya Lipkovich, Quintiles, NC, U.S.
Michael O’Kelly, Quintiles, Dublin, Ireland

ABSTRACT

Multiple imputation (MI) is a methodology for dealing with missing data that has been steadily gaining wide usage in clinical trials. Various methods have been developed and are readily available in SAS PROC MI for multiple imputation of both continuous and categorical variables. MI produces multiple copies of the original dataset, where missing data are filled in with values that differ slightly between imputed datasets. Each of these datasets is then analyzed using a standard statistical method for complete data, and the results from all imputed datasets are combined (pooled) for overall inference using Rubin’s rules which account for the uncertainty associated with imputed values. Rubin’s pooling methodology is very general and is essentially the same no matter what kind of statistic is estimated at the analysis stage for each imputed dataset. However, the combination rules assume that the estimates are asymptotically normally distributed, which may not always be the case. For example, the Cochran-Mantel-Haenszel (CMH) test and the Mantel-Haenszel (MH) estimate of the common odds ratio are often used in analysis of categorical data, and they produce statistics that are not normally distributed. In this case, normalizing transformations need to be applied to the statistics estimated from each imputed dataset before the Rubin’s combination rules can be applied. In this paper, we show how this can be done for the two aforementioned statistics and explore some operating characteristics of the significance tests based on the applied normalizing transformations. We also show how to obtain combined estimates of binomial proportions and their difference between treatment arms.

INTRODUCTION

Multiple imputation (MI) is a methodology introduced by Rubin (1987) for analysis of data where some values that were planned to be collected are missing. In recent years, the problem of missing data in clinical trials received much attention from statisticians and regulatory authorities, which led to a shift in the type of methodologies that are typically used to deal with this problem. In the past, relatively simple approaches, especially single imputation methods were the most popular. For continuous variables, for example, such methods as last observation carried forward (LOCF) or baseline observation carried forward (BOCF) were routinely used. However, a more recent research in this area and the regulatory guidelines from the European Medicines Agency (EMA) (2010) and the FDA-commissioned panel from the National Research Council (NRC) (2010) in US pointed out several important shortcomings that can be encountered when using these methods. One of the concerns is the fact that single imputation methods do not account for the uncertainty associated with missing data and treat single imputation values as if they were real at the analysis stage. This may lead to underestimation of the standard errors associated with estimates of various statistics computed from data. Another problem with these approaches is that, contrary to some beliefs that were quite wide-spread in the past, these methods can bias the analysis in favor of the experimental treatment to an important degree, depending on certain characteristics and patterns of missingness in the clinical study.

Multiple imputation deals directly with the first issue of accounting for the uncertainty of missing data. It does so by introducing in the analysis multiple (but in a sense plausible) values for each missing item and accounting for the variability of these imputed values in the analysis of filled-in data. MI can also be less biased in favor of the experimental treatment under certain assumptions.

Similar to the wide-spread use of single imputation methods for continuous variables in the past, binary and categorical outcomes have been dealt with in a similar way when it came to missing data. For example, for the analysis of clinical trials with a binary outcome, which often represents a status of responder or non-responder to treatment, all cases of missing values were often imputed as non-responders. This is, in a sense, equivalent to the BOCF imputation for continuous variables. In BOCF, subjects are assumed to exhibit the same stage/severity of disease or symptoms at the missing primary time point (typically end of treatment or study) as at baseline, thus assumed not to respond to treatment. Imputing missing binary outcomes to the category of “non-response to treatment” in a deterministic way for all subjects with missing data can thus be expected to present the same problematic issues as does the BOCF, i.e., underestimation of uncertainty and potential bias in favor of experimental treatment. LOCF-type approaches have also been used for binary and categorical data in the past.

Fortunately, multiple imputation can be used not only for continuous variables, but also for binary and categorical ones. This provides for an interesting alternative when there is a concern that single imputation could lead to important bias, and provides a principled way of accounting for uncertainty associated with imputations.

In SAS, PROC MI provides functionality for imputing binary or categorical variables (SAS User's Guide, 2011), of which imputation based on a logistic regression model is probably the most useful in the context of clinical trials. Once a binary or categorical variable is imputed using MI, multiple datasets are created where observed values are the same across all datasets, but imputed values differ. These multiple datasets should then be analyzed using standard methods that would have been chosen should the data have been complete in the first place. Then the results from these multiple datasets are combined (pooled) for overall inference in a way that accounts for the variability between imputations. SAS PROC MIANALYZE provides functionality for combining results from multiple datasets (SAS User's Guide, 2011) which can be readily used after performing a wide range of complete-data analyses.

However, for some types of complete-data analyses, including those for categorical and binary data that are often used in clinical trials, additional manipulations may need to be performed before the functionality of PROC MIANALYZE can be invoked. This is because Rubin's rules (Rubin, 1987) for combining results from multiple imputed datasets implemented by this procedure are based on the assumption that the statistics estimated from each imputed dataset are normally distributed. Many estimates (e.g., means and regression coefficients) are approximately normally distributed, while others, such as correlation coefficients, odds ratios, hazard ratios, relative risks, etc. are not. In this case, a normalizing transformation can be first applied to the estimated statistics, and then the Rubin's combination rules can be applied to the transformed values.

Van Buuren (2012) suggests some transformations that can be applied to several types of estimated statistics (see Table 1 for a partial reproduction of a summary table from Van Buuren's book). He also discusses the methodology for carrying out a multivariate Wald test, likelihood ratio test, chi-square test, and some custom hypothesis tests for model parameters on multiply imputed data, but notes that the last two methods - chi-square test (Rubin, 1997; Li *et al*., 1991) and custom hypothesis tests - may not be very reliable and not enough experience using them in practice is available yet.

Statistic	Transformation
Correlation	Fisher z
Odds ratio	Logarithm
Relative risk	Logarithm
Hazard ratio	Logarithm
Explained variance R^2	Fisher z on root
Survival probabilities	Complementary log-log
Survival distribution	Logarithm

Table 1. Suggested transformations toward normality for various types of statistics. (Partially reproduced from Van Buuren (2012), Table 6.1, p.156.)

In this paper, we discuss normalizing transformations that can be used in order to combine the results of Cochran-Mantel-Haenszel (CMH) test and the odds ratios (from logistic regression or Mantel-Haenszel (MH) estimate of the common odds ratio) based on multiply imputed data, as well as how to obtain combined estimates of binomial proportions and their difference between treatment arms. For the odds ratios we are using the logarithmic transformation. For the CMH test (which is based on a chi-square distributed statistic), in addition to the procedure of Rubin (1987) and Li *et al.* (1991), we use another method where we apply a Wilson-Hilferty transformation to normalize a chi-square distributed statistic. We compare operating characteristics of these approaches using a simulation study.

A detailed background on the workings of MI is beyond the scope of this paper. We provide a very general and high-level discussion on how MI-based analyses can be carried out in SAS and provide some examples of the SAS code to do so. For a detailed treatment of the underlying methodology, we refer readers to Rubin (1987), SAS User's Guide (2011), Carpenter and Kenward (2013), and Van Buuren (2012). The main focus of the paper and examples provided herein is directed to specific steps that need to be implemented in order to obtain overall inferences from analyses that yield non-normally distributed statistics, in particular several analyses of categorical variables mentioned above.

EXAMPLE DATASET

Analysis in this paper will be illustrated using an example dataset, *datain*, with the following variables:

- *subjid* – subject identification number;
- *trt* – treatment arm (0=control and 1=experimental);
- *resp_1, resp_2, resp_3* – binary variables representing response to treatment (0=responder; 1=non-responder) at post-baseline study visits 1, 2, and 3 respectively;
- *score_0* – a continuous baseline score;
- *score_0c* – a baseline score category (1=low; 2= high).

Table 2 summarizes percent of subjects discontinued from the study prior to each study visit and study completers. Some subjects in this dataset discontinued soon after the start of treatment and prior to the first post-baseline visit (14% in the placebo arm and 2% in the experimental arm). These subjects are included in the analysis. Proportion of study completers is somewhat larger in the experimental arm compared to placebo (84% vs. 77%). In this paper, we assume that the input dataset has a monotone pattern of missingness and there are no subjects that missed intermediate visits.

	Discontinued Subjects (Cumulative)	
Visit	**Placebo Arm**	**Experimental Treatment Arm**
1	14%	2%
2	20%	9%
3	23%	16%
	Study Completers	
	77%	84%

Table 2. Percent of subjects discontinuing from the study and study completers

Table 3 shows percentage of responders and non-responders at visit 3 in each treatment arm first based on study completers, and then based on all study subjects if all dropouts are considered to be non-responders (a common single imputation approach used in the past).

	Placebo Arm	**Experimental Treatment Arm**
	Study Completers (Observed Cases)	
Non-responders	82%	68%
Responders	18%	32%
	All Subjects, Dropouts Considered Non-responders	
Non-responders	86%	73%
Responders	14%	27%

Table 3. Percent responders and non-responders at study visit 3 by treatment arm

In subsequent sections we will show how this dataset can be imputed using multiple imputation and then present the results of analysis based on multiply imputed data vs. single imputation (all dropouts as non-responders).

MULTIPLE IMPUTATION IN SAS

Analysis with multiple imputation is generally carried out in three steps:

1. **Imputation**: missing data are filled in using *M* different sets of values which produces *M* imputed datasets. This step can be carried out in SAS using PROC MI.

2. **Analysis**: each of the *M* imputed datasets is analyzed separately using any method that would have been chosen had the data been complete. This step can be implemented using any analytical procedure in SAS, e.g., PROC GLM, PROC MIXED, PROC LOGITIC, PROC FREQ, etc.
3. **Pooling**: analysis results from *M* imputed datasets obtained from step 2 are combined into one overall result. This step can be carried out using SAS PROC MIANALYZE.

SAS procedure PROC MI offers several methods for imputation of both continuous and categorical variables (SAS User's Guide, 2011). The choice of method to use depends, among other things, on whether the missingness pattern is monotone or not. Dataset from a clinical trial will have a monotone missingness pattern if missing values are always due to early withdrawal from the study. That is, when assessments are missing for a given visit, then they will also be missing on all subsequent visits, because subject discontinued study participation. Non-monotone missingness arises when subjects miss some intermediate visits but remain in the study and have available assessments later on. For continuous variables, there is a good choice of imputation methods for both patterns of missingness, whereas for categorical variables, imputation has been mostly limited to monotone missingness in the past, although SAS version 9.3 provides an experimental version of a new class of imputation methods, Fully Conditional Specifications (FCS), which can be used with either pattern and include methods for imputation of categorical data. Nevertheless, even with earlier versions of SAS, there is a way to deal with non-monotone categorical missing data, namely by using Markov Chain Monte Carlo (MCMC) method for partial imputation of non-monotone missing records while treating categorical variables as if they were continuous and modeling them with a multivariate normal distribution. This is not an optimal approach, but is often acceptable because, most of the time, the amount of non-monotone missing data is very small, and the overall impact of this partial imputation step on the analysis at the final study time-point will be small.

Sometimes, a categorical variable is derived based on some underlying continuous measurements. For example, the status of responder to treatment can be determined based on a threshold for a clinically meaningful change from baseline in a continuous parameter, or as an aggregate indicator of changes from baseline in several parameters. In such cases, it is always preferable to first impute the underlying continuous variable(s) and then perform categorization based on imputed values. This way the analyst has a better selection of available methods for imputing continuous variables, and the accuracy of imputations may be improved. However, this approach is not always applicable, as some endpoints are directly defined on a binary or categorical scale.

In PROC MI, two methods are available for imputation of categorical data: logistic regression and discriminant function method. The former method estimates a logistic regression model for each variable that needs to be imputed based on subjects with available data, and then uses predictions from this model (or, more precisely, from a Bayesian posterior distribution of this model and missing data) to fill in missing values. As with other multiple imputation methods, this process is performed in such a way that values sampled for imputation reflect the uncertainty of the estimated logistic regression model (referred to as the imputation model) (Rubin, 1987). Imputation based on a logistic regression model is available for monotone missing data in SAS version 9.3 and below, as well as for non-monotone missingness as one of the experimental FCS approaches, in SAS version 9.3.

Discriminant function method does not seem to be generally useful in the context of clinical trials because its use is limited to cases where all predictor variables are continuous and satisfy the assumptions of approximate multivariate normality and equality of the within-group covariance matrices. In clinical trials, treatment arm is typically represented by a binary or categorical variable and usually needs to be included as a predictor variable in the model. Also, if a categorical endpoint needs to be imputed, we generally wish to include values of this endpoint from previous time-points as predictors. Because of this, the discriminant function method would have a limited utility for clinical trials, but may be useful in some scenarios where a missing baseline categorical covariate needs to be imputed based on a set of some other continuous baseline characteristics.

The choice of the analysis method in step 2 outlined above is guided by the objectives of the study and does not depend in any way on a specific method used for imputation. Any analysis method that would have been used had the data been complete can be applied at this stage. The same analysis method should be used to analyze each of the *M* imputed datasets.

Similarly, the methodology for pooling the results of analysis obtained in step 2 (Rubin, 1987) does not depend on the imputation method used in step 1. The methodology is very general and is essentially the same no matter what kind of statistic is estimated at the analysis stage (e.g., an estimate of the mean or a regression parameter). However, as previously mentioned, the combination rules developed by Rubin rely on the assumption that the estimated statistics are approximately normally distributed. While this assumption holds for many commonly used statistics, it is not the case for some analyses often performed on categorical data. The focus of this paper is on this aspect, and in the subsequent sections, we show what additional steps need to be undertaken in order to combine the results of such analyses from multiply imputed data. Documentation for PROC MIANALYZE, Example 57.10 in SAS User's Guide (2011) illustrates a normalizing transformation that needs to be applied when combining estimates of the Pearson's correlation coefficient before using Rubin's rules implemented in PROC MIANALYZE. We provide examples of other analyses and transformations, focusing on those that are often used in the analysis of categorical data.

PRELIMINARIES: MULTIPLY IMPUTING CATEGORICAL DATA AND COMBINING NORMALLY DISTRIBUTED PARAMETERS

A dataset used as example in this paper is assumed to contain a binary parameter that represents the response to treatment at each study visit. In this section, we illustrate basic steps for performing multiple imputation of this binary data using SAS functionality. We assume that there is no underlying continuous parameter based on which the binary responder status was determined. Therefore, we will be using an imputation method based on logistic regression for imputing categorical variables. This method is available with the MONOTONE statement of PROC MI as shown in SAS Code Fragment 1. When using MONOTONE LOGISTIC statement, PROC MI sequentially estimates a logistic regression imputation model for each variable *resp_1*, *resp_2*, and *resp_3*, where each model includes treatment (*trt*) and baseline score (*score_0*) as predictors. Imputation model for *resp_2* additionally includes *resp_1* as predictor, and the model for *resp_3* includes both *resp_1* and *resp_2*. In general, with the syntax of the MONOTONE statement shown in SAS Code Fragment 1, when a variable with missing values is imputed, all variables listed to its left in the VAR statement are included as predictors in the imputation model. It is possible to specify different models for each variable by using a different syntax in the MONOTONE statement (SAS User's Guide, 2011). The option NIMPUTE in the PROC MI statement specifies the number of imputed datasets to be generated. The output dataset *datain_mi* will contain 500 copies of the original dataset, with the observed values being the same across all datasets, and with imputed values varying from one dataset to another. These multiple copies will be identified by a new variable, *_Imputation_,* added to the output dataset by PROC MI.

```
PROC MI DATA=datain OUT=datain_mi SEED=4566765 NIMPUTE=500;
    VAR trt score_0 resp_1 - resp_3;
    CLASS trt resp_1 - resp_3;
    MONOTONE LOGISTIC;
RUN;
```

SAS Code Fragment 1. Multiple imputation of binary response variables using logistic regression

When using logistic regression for imputation, the analyst should be aware of a potential problem of perfect prediction. It may occur if the strata formed by the covariates included in the model form cells in which all available values of a dependent categorical variable are the same (e. g., available binary outcomes are all 0s or all 1s within a cell). This may result in the imputation model generating imputed values that are very different from observed ones (see Carpenter and Kenward (2013) for more details on this issue). In clinical trials, this may be more likely with the imputation of a binary responder status at earlier time-points if achievement of response is not likely at the beginning of the study. Also, this may happen if such covariates as investigator site are included in the model, and there are sites with a relatively small number of subjects, all having the same response at a given time-point. To deal with this potential problem, it is advisable to carry out a preliminary exploratory step by fitting logistic regression models to available data at each time-point, e.g., using PROC LOGISTIC, and carefully examining the resulting model parameters. PROC LOGISTIC will produce a warning of a "quasi complete separation" in this case, and the analyst can subsequently modify the model by excluding or changing certain covariates to avoid this problem. Once the models have been appropriately selected for each time-point, they can be specified in PROC MI by using a separate MONOTONE LOGISTIC statement with a distinct model for each variable.

After imputation is performed, the next step is to analyze the imputed datasets. SAS Code Fragment 2 provides an example of analysis where a logistic regression model is used to estimate the effect of treatment on response at study visit 3 adjusting for baseline score as continuous covariate. The output dataset *datain_mi* from PROC MI is used as input to the analysis procedure PROC LOGISTIC, and because this dataset contains 500 imputed copies of the original dataset, the analysis procedure is invoked with a "BY *_Imputation_*" statement, so that the same analysis is performed within each of the imputed datasets. The ODS output datasets PARAMETERESTIMATES and ODDSRATIOS are saved to capture estimates of the regression coefficients and odds ratios respectively estimated by the analysis model. These ODS datasets will contain a set of estimates for each imputed dataset identified by the variable *_Imputation_* included in each of them.

SAS Code Fragment 2 also shows an invocation of PROC MIANALYZE which is used to combine the results of analyses from PROC LOGISTIC on multiply imputed dataset *datain_mi*. ODS output dataset PARAMETERESTIMATES saved under the name *lgsparms* and containing estimates of the logistic regression coefficients is passed as input to PROC MIANALYZE using PARMS option. This is one of the options used for conveying the analysis results and the information about the structure of the datasets containing them. The PARMS option in the PROC MIANALYZE statement is used to pass a dataset which contains parameter estimates and the associated standard errors. Option CLASSVAR=CLASSVAL included in parentheses indicates to PROC MIANALYZE some additional information about the structure of the input dataset in which the levels for the classification effects are specified. SAS documentation for PROC MIANALYZE includes an extensive set of examples of analyses with many different analytical SAS procedures and the appropriate syntax to pass their results to PROC MIANALYZE.

In the invocation of PROC MIANALYZE in SAS Code Fragment 2, we specify *trt* variable in MODELEFFECTS statement, by which we request an overall (pooled) estimate of the regression coefficient for treatment effect. The output from this procedure will thus provide us a combined regression estimate, its standard error, confidence interval (CI), and p-value from a hypothesis test of the coefficient being equal to 0 (no treatment effect).

```
PROC LOGISTIC DATA=datain_mi;
   CLASS trt(DESC);
   MODEL resp_3(EVENT='1') = score_0 trt ;
   ODS OUTPUT PARAMETERESTIMATES=lgsparms ODDSRATIOS=lgsodds;
   BY _Imputation_;
RUN;

PROC MIANALYZE PARMS(CLASSVAR=CLASSVAL)=lgsparms;
   CLASS trt;
   MODELEFFECTS trt;
   ODS OUTPUT PARAMETERESTIMATES=mian_ lgsparms;
RUN;
```

SAS Code Fragment 2. Analysis of multiply imputed data using logistic regression and pooling estimates of the regression coefficient for treatment effect.

Standard multiple imputation, as is illustrated in this example, operates under a Missing at Random (MAR) assumption about the missingness mechanism. Under MAR, withdrawn subjects are assumed to have the same probability distribution for response to treatment at time points after their study discontinuation as subjects who remained in the study, conditional on baseline and pre-withdrawal data included in the analysis. In other words, discontinued subjects are assumed to have the same probability of response as similar subjects who remained in the study. This is in contrast with the assumption that all dropouts would be non-responders (with probability 1) often done in single imputation analysis. Because completers will generally have a non-zero probability of responding to treatment, the MI imputation will result in more optimistic outcomes being imputed for missing values in both treatment arms. When the proportion of discontinuations is larger in the placebo arm, as is the case in our example dataset, the MI imputation may be more favorable to placebo than the all-non-responder imputation. In this case, the MI imputation is likely to be less biased in favor of the experimental treatment compared to the all-non-responder approach, and could be considered more appropriate as per the regulatory guidance (EMA, 2010; NRC, 2010).

The results from analysis of multiply imputed data as described above, as well as from a single imputation approach imputing all dropouts as non-responders are shown in Table 4. We also present the results of observed cases analysis (no imputation) for reference purposes. Under an MAR assumption, observed cases analysis provides an unbiased estimate of an odds ratio and consequently of the coefficient corresponding to treatment effect in a logistic regression model (Carpenter and Kenward, 2013). As would be expected based on the discussion in the previous paragraph, the MI-based analysis produces a slightly smaller estimate of the treatment effect coefficient and the corresponding p-value compared to the single imputation approach. The MI-based estimate is also closer to the one from observed cases analysis. The treatment effect is statistically significant under all analyses.

	Estimate of regression coefficient for treatment effect (95% CI)	**P-value**
Multiple Imputation	0.41 (0.16, 0.66)	0.0013
Single Imputation	0.46 (0.22, 0.70)	0.0002
Observed Cases	0.38 (0.13, 0.64)	0.0035

Table 4. Estimate of treatment effect in a logistic regression model based on multiple imputation vs. single imputation and observed cases

In SAS Code Fragment 2, the ODS output dataset PARAMETERESTIMATES was passed as is to PROC MIANALYZE. Estimates of regression coefficients are approximately normally distributed and thus Rubin's combination rules implemented by this SAS procedure can be directly applied. The same cannot be said about the estimates of the odds ratio captured in the ODS output dataset ODDSRATIOS as these estimates have a log-normal distribution. We show how to handle this case in the nest section.

POOLING ODDS RATIOS USING LOG TRANSFORMATION

As mentioned above, the estimates of odds ratios follow a log-normal distribution. We can apply a log transformation to normalize these estimates in order to be able to apply Rubin’s combination rules. As previously mentioned, these combination rules take as input estimates of a statistic obtained from multiple imputed datasets as well as standard errors of these estimates, and produce the overall pooled estimate, overall standard error (variance), confidence interval and a p-value from a univariate hypothesis test of the statistic being equal to zero. The first data step in SAS Code Fragment 3 (performed on the ODS output dataset ODDSRATIOS from PROC LOGISTIC saved under name *lgsodds*) contains a log transformation applied to the estimates of the odds ratio for the treatment effect. Standard error of the transformed estimate is obtained from the log-transformed lower and upper confidence limits for the odds ratio estimate. Then the dataset containing the transformed estimates and their standard errors is passed to PROC MIANALYZE as illustrated in the same code fragment. In this case, a different syntax of input is used with PROC MIANALYZE using a DATA option. With this option, the MODELEFFECTS statement contains a name of the variable that represents an estimate of the statistic to be combined, and the STDERR statement contains the name of the variable that represents standard errors of that estimate. The combined results are captured in the ODS dataset PARAMETERESTIMATES. The combined estimate of the odds ratio can then be back-transformed to its original log scale as shown in the last data step of SAS Code Fragment 3, which also computes confidence limits on the log scale using the combined estimate of the standard error for the odds ratio.

```
*** Log-transform odds ratio estimates
    and obtain standard error from confidence intervals ***;
DATA lgsodds_t; SET lgsodds(WHERE=(INDEX(EFFECT,"TRT")));
   log_or_lr_value=LOG(ODDSRATIOEST);
   log_or_lr_se=(LOG(UPPERCL)-LOG(LOWERCL))/(2*1.96);
RUN;

*** Combine transformed estimates;
PROC MIANALYZE DATA=lgsodds_t;
   ODS OUTPUT PARAMETERESTIMATES=mian_lgsodds_t;
   MODELEFFECTS log_or_lr_value;
   STDERR log_or_lr_se;
RUN;

*** Back-transform combined values;
DATA mian_lgsodds_bt; SET mian_lgsodds_t;
   Estimate_back = EXP(ESTIMATE);            *Pooled odds ratio;
   LCL_back=Estimate_back*EXP(-1.96*STDERR); *Pooled lower limit;
   UCL_back=Estimate_back*EXP(+1.96*STDERR); *Pooled upper limit;
RUN;
```

SAS Code Fragment 3. Pooling estimates of the odds ratio obtained from the analysis of multiply imputed data using logistic regression

Table 5 presents the odds ratio estimates and their confidence intervals from the analysis of multiply imputed data as described above (combined estimate), from the analysis using single imputation, and from the observed cases analysis. Once again, as expected, we see that the MI estimate is slightly smaller than that from single imputation, and close to the one from observed cases, and none of the confidence intervals cover 1, representing an odds ratio for treatment effect that is statistically significantly different from 1.

	Estimate of odds ratio (experimental treatment vs. placebo)	95% Confidence Interval
Multiple Imputation	2.27	1.38, 3.74
Single Imputation	2.49	1.54, 4.04
Observed Cases	2.15	1.29, 3.60

Table 5. Estimate of odds ratio for experimental treatment vs. placebo obtained from a logistic regression model based on multiple imputation vs. single imputation and observed cases

Our example dataset is based on a binary responder variable, but logistic regression can also be used for variables with multiple categorical levels, and the same principles for handling the resulting statistics (regression coefficients and odds ratios) would apply in that case by selecting the estimates for appropriate categorical levels of interest.

For binary outcomes, odds ratio can also be obtained using the Mantel-Haenszel estimate of the common odds ratio (Mantel & Haenszel, 1959; Agresti, 2002) which in SAS can be computed by PROC FREQ for adjusted 2×2 tables (e.g., 2 levels of treatment arm and a binary response variable, as in our example, adjusted for a categorical stratification variable). The transformation needed in this case would be exactly the same as described above. SAS Code Fragment 4 illustrates this with PROC FREQ used to perform Mantel-Haenszel analysis with baseline score category as stratification factor and subsequent log-transformation of the common odds ratio estimates in the ODS output dataset COMMONRELRISKS. Combining the transformed estimates with PROC MIANALYZE and back-transformation steps would be identical to those in SAS Code Fragment 3.

```
*** Obtain Mantel-Haenszel estimate of the common odds ratio adjusted for
    baseline score category ***;
PROC FREQ DATA=datain_mi;
   TABLES score_0c*trt*resp_3 / CMH;
   ODS OUTPUT COMMONRELRISKS=comrrout;
   BY _Imputation_;
RUN;

*** Log-transform odds ratio estimates
    and obtain standard error from confidence intervals ***;
DATA ormh_t; SET comrrout(WHERE=(StudyType="Case-Control"));
   log_or_mh_value=log(VALUE);
   log_or_mh_se=(log(UPPERCL)-log(LOWERCL))/(2*1.96);
RUN;
```

SAS Code Fragment 4. Transforming estimates of the Mantel-Haenszel common odds ratio obtained from analysis of multiply imputed data

Table 6 contains Mantel-Haenszel estimates of the common odds ratios from multiply imputed data vs. single imputation, as well as for the observed cases analysis. These estimates are close to those from logistic regression.

	Estimate of odds ratio (experimental treatment vs. placebo)	**95% Confidence Interval**
Multiple Imputation	2.21	1.36, 3.58
Single Imputation	2.46	1.52, 3.98
Observed Cases	2.20	1.33, 3.64

Table 6. Estimates of the Mantel-Haenszel common odds ratio for experimental treatment vs. placebo based on multiple imputation vs. single imputation and observed cases

POOLING RESULTS OF THE COCHRANE-MANTEL-HAENSZEL TEST USING WILSON-HILFERTY TRANSFORMATION

Cochran-Mantel-Haenszel test (Landis *et al.,* 1978) is often used in the analysis of clinical trials for a complete-data analysis of the relationship between two categorical variables (e.g., treatment group and response to treatment) after controlling for one or more stratification variables (e.g., baseline disease severity) in a multi-way table. The CMH general association statistic, under the null hypothesis of no association, has an asymptotic chi-square distribution with $(C_1 - 1)(C_2 - 1)$ degrees of freedom where C_1 and C_2 represent the number of categories assumed by each of the two categorical variables. The chi-square distribution is highly skewed for smaller degrees of freedom, and thus obtaining a combined result of the CMH test from multiply-imputed data requires a transformation that would normalize the CMH statistic. For example, the Wilson-Hilferty transformation (Wilson & Hilferty, 1931; Goria, 1992) can be used for this purpose:

$$wh\_cmh^{(m)} = \sqrt[3]{cmh^{(m)}/df} \qquad (1)$$

where $cmh^{(m)}$ is the CMH statistic computed from the m^{th} imputed dataset, df is the number of degrees of freedom associated with the CMH statistic, and $wh\_cmh^{(m)}$ is the transformed value. The transformed statistic is approximately normally distributed with mean $1 - 2/(9 \times df)$ and variance $2/(9 \times df)$ under the null hypothesis.

We can standardize this transformed statistic in (1) to obtain a variable that is normally distributed with mean 0 and variance 1:

$$st_{wh\_cmh}{}^{(m)} = \frac{\sqrt[3]{\frac{cmh^{(m)}}{df}} - \left(1 - \frac{2}{9 \times df}\right)}{\sqrt[2]{\frac{2}{9 \times df}}} \qquad (2)$$

This transformed statistic can now be passed to PROC MIANALYZE in order to perform a combined CMH test.

SAS Code Fragment 5 contains an invocation of PROC FREQ to request the CMH test using the CMH option in the TABLES statement, with the results captured in the ODS output dataset CMH. A subsequent data step applies the Wilson-Hilferty transformation as described in equation (2) and then passes the transformed values to PROC MIANALYZE using the same syntax as for the odds ratio. Finally, a p-value for the combined CMH test can be obtained as the upper-tailed p-value from the normal test produced by PROC MIANALYZE on the transformed statistic. This is done in the last DATA step of the SAS Code Fragment 5.

```
*** Perform CMH test;
PROC FREQ DATA=datain_mi;
   TABLES score_0c*trt*resp_3 / CMH;
   ODS OUTPUT CMH=cmh;
   BY _Imputation_;
RUN;

*** Apply Wilson-Hilferty transformation to the CMH statistic and
    standardize the resulting normal variable;
DATA cmh_wh; SET cmh(WHERE=(AltHypothesis="General Association"));
   cmh_value_wh=((VALUE/DF)**(1/3) - (1-2/(9*DF)))/SQRT(2/(9*DF));
   cmh_sterr_wh = 1.0;
RUN;

*** Combine results;
PROC MIANALYZE DATA=cmh_wh;
   ODS OUTPUT PARAMETERESTIMATES=mian_cmh_wh;
   MODELEFFECTS cmh_value_wh;
   STDERR cmh_sterr_wh;
RUN;

*** Compute one-sided p-value;
DATA mian_cmh_wh_p; SET mian_cmh_wh;
   IF tValue > 0 THEN Probt_upper = Probt/2;
   ELSE Probt_upper = 1-Probt/2;
RUN;
```

SAS Code Fragment 5. Pooling estimates of the CMH statistic from analysis of multiply imputed data and obtaining an overall p-value for the CMH test

Table 7 shows the p-values obtained from the CMH test for general association between treatment arm and responder status adjusting for baseline score category from multiply imputed data, from single imputation and from observed cases. The p-values from this test are close to those for the regression coefficient of treatment effect from the logistic regression model (see Table 3), with CMH test on multiply imputed data being somewhat more conservative than the test on single imputation data.

	P-value
Multiple Imputation	0.0011

Single Imputation	0.0002
Observed Cases	0.0021

Table 7. Overall p-values for CMH test of general association based on multiple imputation vs. single imputation and observed cases

To illustrate how the Wilson-Hilferty transformation affects the hypothesis test, Figure 1 presents a scatter plot of p-values obtained when comparing a range of untransformed chi-square statistics to the chi-square distribution with 1 degree of freedom (on the x-axis) vs. those obtained from comparing the transformed statistics to the normal distribution (y-axis). Left panel of Figure 1 shows that the scatter points (represented by circles) follow a line close to identity on most of the range of the p-values, meaning that the hypothesis test on a transformed statistic would give approximately the same p-value as a test on the untransformed chi-square statistic. Only towards the high end of the p-values (>0.8), the transformed test would provide smaller p-values which, however, would not alter the conclusion of statistical significance.

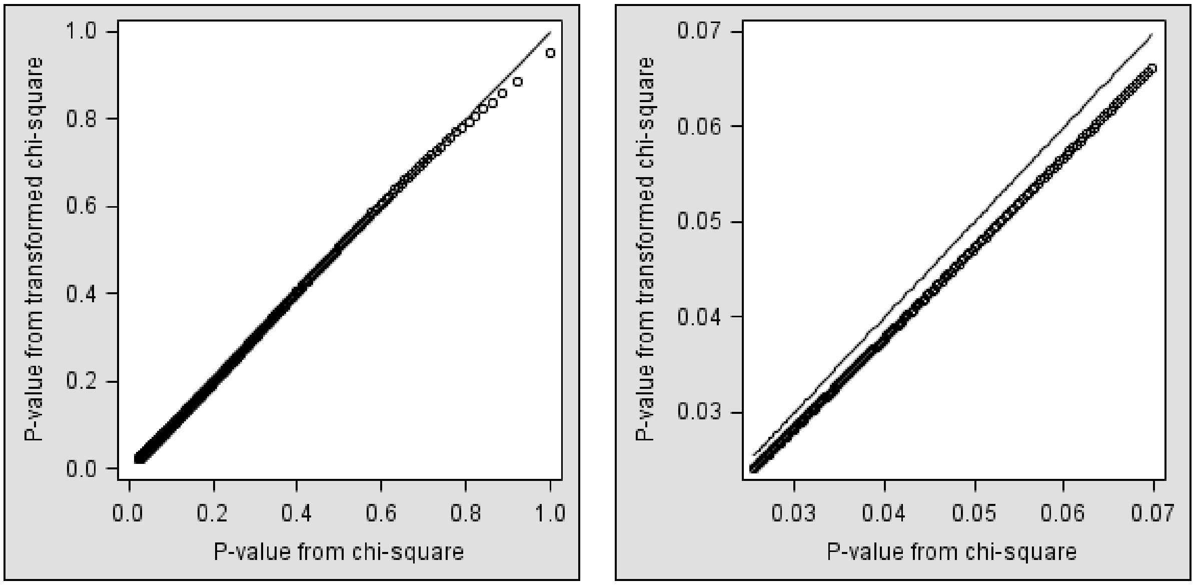

Figure 1. Scatter plot of p-values from an untransformed chi-square statistic compared to the chi-square distribution with 1 degree of freedom vs. p-values from the Wilson-Hilferty transformed chi-square statistic compared to the normal distribution

If we zoom in at the range of p-values close to the statistical significance level of 0.05 as presented on the right panel of Figure 1, we can see that the p-values from the transformed statistic are slightly lower. Table 8 provides a list of p-values which would lead to a conclusion of no statistical significance (or border-line) based on the untransformed chi-square test, whereas the p-values from a transformed statistic would be slightly below the significance level of 0.05. In all other cases, the conclusion regarding the statistical significance would be the same. This small difference should be taken into consideration when interpreting border-line significant findings. Multiply-imputed data provide a more conservative estimate of the treatment effect compared to single imputation, in general, and thus using multiple imputation with this transformation is still likely to be more conservative than using single imputation.

It should be noted that the distribution for the Wilson-Hilferty transformed statistic would not be the same under the alternative hypothesis because the underlying CMH statistic would have a non-central chi-square distribution with an unknown non-centrality parameter. This would have an impact if Rubin’s combination rules were applied to obtain combined estimates of confidence intervals, but the combined CMH hypothesis test should still be appropriate under this transformation. More discussion is provided below based on the results of a simulation study using different tests.

P-value from chi-square statistic	P-value from a Wilson-Hilferty transformed chi-square statistic

0.0500	0.0473
0.0503	0.0476
0.0506	0.0478
0.0509	0.0481
0.0513	0.0484
0.0516	0.0487
0.0519	0.0490
0.0522	0.0493
0.0525	0.0496
0.0528	0.0499

Table 8. Range of p-values from an untransformed chi-square statistic compared to the chi-square distribution with 1 degree of freedom vs. p-values from the Wilson-Hilferty transformed chi-square statistic compared to the normal distribution where they disagree as to the statistical significance at the 0.05 level

POOLING CHI-SQUARE STATISTICS USING PROCEDURE OF RUBIN (1987) AND LI *ET AL.* (1991)

In this section, we describe an alternative procedure for pooling chi-square distributed statistics that was proposed by Rubin (1987) and further investigated by Li *et al.* (1991). Denote by χ_m^2 chi-square distributed statistics with *k* degrees of freedom estimated in each of the *m=1,…, M* imputed datasets. A pooled test statistic can be obtained as follows:

$$D_x = \frac{\dfrac{\bar{\chi}^2}{k} - \dfrac{M+1}{M-1}\bar{r}_x}{1+\bar{r}_x}\text{, where} \tag{3}$$

$$\bar{\chi}^2 = \frac{1}{M}\sum_{m=1}^{M}\chi_m^2\,;\ \bar{r}_x = (1+\frac{1}{M})\frac{1}{M-1}\sum_{m=1}^{M}\left(\sqrt{\chi_m^2} - M^{-1}\sum_{m=1}^{M}\sqrt{\chi_m^2}\right)^2$$

The pooled p-value for the hypothesis test based on D_x can be obtained using *F* distribution with *k* and ν_x as numerator and denominator degrees of freedom, respectively, as follows:

$$P_x = \Pr[F_{k,\nu_x} > D_x];\nu_x = k^{-3/M}(M-1)(1+\frac{1}{\bar{r}_x})^2 \tag{4}$$

A macro implementing the steps described by equations (3) and (4) is provided in SAS Code Fragment 6, where it is assumed that the estimated statistics from the *M* imputed datasets are saved in a dataset with a variable *chsq_value*, and this dataset is passed to the macro using the *datain* argument. In the case of unadjusted analysis of a 2×2 table, the number of degrees of freedom associated with each chi-square distributed statistic is 1, but this value can vary for other analyses and can be passed to the macro using the *df* argument.

```
*** Implement pooling procedure described in equations (3) and (4);
%MACRO computePooledCh(datain,dataout,df=1);
    PROC IML;
      USE &datain ;
      READ ALL VAR {chsq_value} INTO chval;
      df=&df;
      m=NROW(chval);
      cvalroot_m = sum(chval##0.5)/m;
      cval_m = SUM(chval)/m;
      a=(chval##0.5-j(m,1,1)*cvalroot_m)##2;
      rx = sum(a)*(1+1/m)/(m-1);
      Dx=(cval_m/df - (m+1)/(m-1)*rx)/(1+rx);
      df_den=(df**(-3/m))*(m-1)*(1+1/rx)**2;
      Pval=1-CDF("F",Dx,df,df_den);
      CREATE &dataout FROM Pval[colname={"PvalPooledCh"}];
      APPEND FROM Pval;
    RUN; QUIT;
%MEND;
```

SAS Code Fragment 6. Macro to implement a procedure by Rubin (1987) and Li et al. (1991) for pooling chi-square distributed statistics

We will use this method in the simulation study described below and compare its operating characteristics to the method based on Wilson-Hilferty transformation.

POOLING ESTIMATES OF BINOMIAL PROPORTIONS IN EACH TREATMENT ARM AND THE DIFFERENCE BETWEEN PROPORTIONS

In this section, we show how to combine the estimates of binomial proportions of responders in each treatment arm, and the difference between these proportions. For the proportions of responders in each treatment arm, the proportion estimates and their asymptotic standard errors provided by PROC FREQ (captured in ODS output dataset BINOMIALPROP) for each imputed dataset can be passed directly to PROC MIANALYZE, as shown in SAS Code Fragment 7. No transformation is needed; the data step after the invocation of PROC FREQ simply aligns the proportion estimates and their standard errors on one record for each imputed dataset, and the resulting dataset is then passed to PROC MIANALYZE.

```
*** Estimate proportions of responders in each treatment arm;
PROC FREQ DATA=datain_mi;
   TABLES resp_3 / cl binomial(level=2);
   BY _Imputation_ trt;
   ODS OUTPUT BINOMIALPROP=prop;
RUN;


*** From ODS output dataset BINOMIALPROP, create a dataset
    containing estimated proportion of responders in each
    treatment arm and their standard errors;
DATA prop_trt;
   MERGE
     prop(WHERE=(Label1="Proportion")
          KEEP=_Imputation_ trt nValue1 Label1
          RENAME=(nValue1=prop))
     prop(WHERE=(Label1="ASE")
          KEEP=_Imputation_ trt nValue1 Label1
          RENAME=(nValue1=prop_se));
   BY _Imputation_ trt;
RUN;
```

```
*** Combine proportion estimates;
PROC SORT DATA=prop_trt; BY trt _Imputation_; RUN;
PROC MIANALYZE DATA=prop_trt;
   MODELEFFECTS prop;
   STDERR prop_se;
   BY trt;
   ODS OUTPUT PARAMETERESTIMATES=mian_prop_trt;
RUN;
```

SAS Code Fragment 7. Pooling estimates of the binomial proportions from analysis of multiply imputed data

For the difference between proportions in two arms, the standard error of the estimated difference is computed as the square root of the sum of squared standard errors for each proportion. This standard error is then passed to PROC MIANALYZED along with the estimated difference in proportions, as shown in SAS Code Fragment 8.

```
*** Compute estimates of the difference in proportions of
    responders between treatment arms and their standard errors;
DATA prop_diff;
   MERGE prop_trt(WHERE=(trt=0) RENAME=(prop=p1 prop_se=se1))
         prop_trt(WHERE=(trt=1) RENAME=(prop=p2 prop_se=se2));
   BY _Imputation_;

   prop_diff = (p2-p1);
   se_diff = sqrt(se1*se1 + se2*se2);
RUN;

*** Combine estimates of the proportion differences;
PROC MIANALYZE DATA=prop_diff;
   MODELEFFECTS prop_diff;
   STDERR se_diff;
   ODS OUTPUT PARAMETERESTIMATES=mian_prop_diff;
RUN;
```

SAS Code Fragment 8. Pooling estimates of the difference between binomial proportions from analysis of multiply imputed data.

Table 9 summarizes the estimates of the proportion of responders in each treatment arm and their difference based on multiply imputed data vs. single imputation and observed cases. Multiple and single imputation estimates for the difference between proportions in this case are almost the same, but the p-value from the test of this difference being greater than zero is somewhat larger from the multiply imputed data. These p-values agree with those from other analyses discussed above. The estimated proportions of responders in each treatment arm are quite different based on multiple imputation compared to single imputation. With multiple imputation, there are more responders estimated in each treatment arm, which would be expected because multiple imputation models response based on study completers and thus at least some dropouts are likely to be similar to completers and have a non-zero chance of responding, compared to a zero probability of response in our single imputation approach.

	Proportion of responders in placebo arm (95% CI)	**Proportion of responders in experimental arm (95% CI)**	**Difference in proportions of responders in experimental and placebo arms (95% CI)**
Multiple Imputation	0.19 (0.14, 0.24)	0.32 (0.26, 0.38)	0.13 (0.05, 0.21) p-value=0.0012
Single Imputation	0.14 (0.10, 0.19)	0.27 (0.22, 0.33)	0.13 (0.06, 0.21) p-value=0.0002
Observed Cases	0.18 (0.13, 0.24)	0.32 (0.26, 0.39)	0.14 (0.05, 0.23) p-value=0.0007

Table 9. Estimates of the proportion of responders in each treatment arm and their difference based on multiply imputed data vs. single imputation and observed cases

SIMULATION STUDY

We conducted a small simulation study in order to examine and compare operating characteristics (power and Type I error rate) of three methods of analysis and pooling with multiply imputed binary data: (1) MH estimate of the common odds ratio with a logarithmic transformation (MHOR-LT); (2) the MH test with Wilson-Hilferty transformation (MH-WHT); and (3) the MH test with the chi-square pooling procedure (MH-CHP) of Rubin (1987) and Li *et al.* (1991). Note that in the context of 2×2 table analysis (binary responder status for two treatment arms), the MH test is equivalent to the unadjusted CMH test discussed earlier. The Wilson-Hilferty transformation and the other chi-square pooling procedure can also be applied to a more general setting using stratified CMH test for $C_1 \times C_2$ tables at the analysis stage as previously shown.

Our simulation study mimics data that could arise from a clinical trial. To generate simulated data, we assumed that the response to treatment is defined based on some underlying continuous endpoint and a pre-specified responder cut-off value, i.e., 50% improvement from baseline in the continuous value. We first simulated data for this underlying continuous variable across one baseline and 3 post-baseline evaluation visits using a multivariate normal distribution. The covariance structure was chosen to reflect a realistic situation where the within-subject correlation decreases as the measurements get farther apart in time. The underlying mean structure was calibrated so as to match pre-specified rates of response (based on a 50% improvement from baseline definition). Two levels for the placebo arm response rates were simulated: a lower rate of 20%, and a higher rate of 40%. For the experimental arm subjects, we considered two corresponding null scenarios where the rates for the experimental arm were exactly the same as for placebo subjects, and two alternative scenarios where the response rates in the experimental arm were calibrated so as to ensure an approximately 80% power when testing for the difference between proportions assuming no discontinuations occurred. The analysis was intended to be performed on the resulting 2×2 table - binary responder status for two treatment arms – without any stratification factor.

We assumed a rather simplistic MAR model of missingness, where only subjects whose outcome worsened by some amount equal to or greater than some cut-off value were “eligible” to drop out at post-baseline visits 2 or 3. Once a subject reached this “drop-out condition”, s/he was assumed to discontinue with the probability γ. While holding γ=0.5, we calibrated the drop-out eligibility cut-off so as to ensure the desired percentage of the dropout in the placebo arm. The dropout rate in experimental arm would then be driven by a combination of 2 factors: the rate of dropout in the placebo arm and the assumed treatment effect size (odds ratio).

Table 10 summarizes parameters corresponding to different simulation scenarios studied (more details are provided in the Appendix).

The primary focus of this simulation study was to evaluate the three methods of analysis and pooling the results from multiply-imputed binary data. Overall operating characteristics can also be affected by the imputation model, and in order to take this into account, we used two methods of imputation: (1) imputing the binary response variable directly using sequential logistic regression; (2) imputing the underlying continuous variable using ordinary sequential linear regression and then computing the binary responder status based on the observed and imputed continues values. In both cases, the imputation models were similar in the sense that they included a baseline value and post-baseline outcomes from all previous time-points within each treatment arm in each of the sequential regression models.

Data simulated under null or alternative hypothesis	Dropout rates in placebo arm	Responder probability for placebo	Responder probability for experimental	Comments
Null	10%, 20%, 30%, 40%	0.4	0.4	To evaluate Type I error rate

Alternative	10%, 20%, 30%, 40%	0.4	0.609
			OR=2.3363
			LOR=0.8486
				Corresponds to power of 0.80 for the chi-squared test on proportions in the complete data (N=100 per arm)

Null	10%, 20%, 30%, 40%	0.2	0.2	To evaluate type I error rate

Alternative	10%, 20%, 30%, 40%	0.2	0.391
			OR=2.568
			LOR=0.943127
				Corresponds to power of 0.80 for the chi-squared test on proportions in the complete data (N=100 per arm)

Table 10. Summary of simulation scenarios

The results of the simulation study (based on 1000 simulated datasets and 100 imputations) are presented in Tables 11 and 12 which report Type I error rates and statistical power respectively. In addition to the results from imputation and analysis/pooling methods mentioned above, these tables also report the results from the analysis of observed cases estimating MH common odds ratio. This analysis can be considered as a good benchmark in the current simulation setting, because the available cases analysis provides an unbiased estimate of the odds ratio in the case of MAR mechanism (Carpenter and Kenward, 2013).

We can see from Table 11 that all 3 analysis methods (MHOR-LT, MH-WHT, and MH-CHP) applied to the multiply-imputed data maintained the nominal Type I error rates (<0.05) regardless of the imputation model used, with the exception of the MH-CHP method with 10% dropout rate which resulted in slightly inflated error rates. The error rates were quite similar between the three methods, as well as similar to those from available cases analysis. MHOR-LT method was slightly more conservative than the others. The MH-WHT method had slightly higher rates (but within nominal level) compared to the other two analysis/pooling methods for simulation scenarios with higher dropout rates when imputations were performed on a binary scale. Overall, the differences between methods are quite small and may be due, to some extent, to the simulation error.

Responder probability for placebo group	Dropout rate in placebo arm	Imputation and analysis methods						Observed cases analysis
		Imputing on a binary scale			Imputing on a continuous scale			
		MHOR-LT	MH-WHT	MH-CHP	MHOR-LT	MH-WHT	MH-CHP	MHOR
20%	10%	0.038	0.038	0.036	0.036	0.039	0.044	0.038
	20%	0.026	0.029	0.027	0.036	0.042	0.048	0.043
	30%	0.036	0.040	0.036	0.039	0.047	0.047	0.046
	40%	0.033	0.045	0.033	0.027	0.034	0.031	0.036
40%	10%	0.046	0.048	0.055	0.044	0.046	0.050	0.050
	20%	0.043	0.049	0.047	0.027	0.034	0.031	0.036
	30%	0.034	0.039	0.033	0.035	0.039	0.040	0.048
	40%	0.034	0.038	0.028	0.033	0.036	0.030	0.051

MHOR-LT = Mantel-Haenzel estimate of common odds ratio with logarithm transformation; MH-WHT = Mantel-Haenzel test with Wilson-Hilferty transformation; MH-CHP = Mantel-Haenzel test with chi-square pooling procedure of Rubin (1987) and Li et al. (1991); MHOR = Mantel-Haenzel estimate of common odds ratio on observed cases data.

Table 11. Type I error rates from a simulation study

The differences among the 3 analysis/pooling methods applied to multiply imputed data in terms of statistical power are also rather small: under alternative scenarios MH-CHP and MH-WHT appear to slightly outperform MHOR-LT in the context of both imputation methods, with MH-WHT gaining a small advantage over MH-CHP as the dropout rate increases especially when imputation is done on the binary scale.

While the differences between the analysis/pooling methods applied to imputed data are very small, the differences in power for different imputation approaches are more notable. Firstly, the observed cases analysis, while being unbiased in the present context, shows losses of power that range between 7% and 15% depending on the dropout rate and placebo responder probability. Imputing data on a continuous scale allows us to regain up to 5-6% of that power loss. Imputing data directly on the binary scale, however, results in more drastic power loses compared to the observed cases – up to 15% in some cases. This result is not surprising as coarsening predictors in the imputation model would be expected to result in some loss of information, which translates into lower power.

Responder probability for placebo group	**Dropout rate in placebo group**	**Imputation and analysis methods**						**Observed cases analysis**
		Imputing on a binary scale			**Imputing on a continuous scale**			
		MHOR-LT	**MH-WHT**	**MH-CHP**	**MHOR-LT**	**MH-WHT**	**MH-CHP**	**MHOR**
20%	10%	0.622	0.644	0.652	0.751	0.764	0.790	0.730
	20%	0.538	0.559	0.557	0.712	0.718	0.742	0.685
	30%	0.462	0.499	0.454	0.660	0.682	0.688	0.630
	40%	0.387	0.425	0.357	0.588	0.623	0.597	0.571
40%	10%	0.663	0.675	0.695	0.742	0.758	0.770	0.721
	20%	0.571	0.591	0.571	0.695	0.707	0.728	0.662
	30%	0.472	0.495	0.464	0.633	0.651	0.650	0.621
	40%	0.378	0.405	0.356	0.576	0.596	0.576	0.558

MHOR-LT = Mantel-Haenzel estimate of common odds ratio with logarithm transformation; MH-WHT = Mantel-Haenzel test with Wilson-Hilferty transformation; MH-CHP = Mantel-Haenzel test with chi-square pooling procedure of Rubin (1987) and Li et al. (1991); MHOR = Mantel-Haenzel estimate of common odds ratio on observed cases data.

Table 12. Statistical power from a simulation study

DISCUSSION AND CONCLUSION

In this paper, we focused on the analysis of multiply imputed categorical data, and in particular on how to combine the results of categorical analyses from MI for overall inference. Rubin's combination rules rely on the assumption of approximately normal distribution of the statistics estimated in each of the imputed datasets, which is not always the case with categorical analyses. We showed how normalizing transformations can be used on the estimated statistics prior to applying Rubin's rules to combine the results of Cochran-Mantel-Haenszel test, log odds ratio and Mantel-Haenszel estimate of the common odds ratio, as well as the estimates of binomial proportions and their difference between treatment arms. These analyses are quite common in clinical trials, and we hope that the examples presented in this paper will facilitate application of multiple imputation to categorical data. Multiple imputation makes different assumptions about unobserved outcomes of the discontinued subjects compared to common single imputation methods, and in some situations, can be regarded to be more plausible, and/or less biased in favor of the experimental treatment. Multiple imputation can be used under both MAR and MNAR assumptions, and once the data are appropriately imputed, the analysis and pooling methods are the same.

A small simulation study that we conducted suggests that for a hypothesis test related to treatment effect, both a Wilson-Hilferty transformation and a chi-square pooling procedure by Rubin (1987) and Li *et al.* (1991) can be successfully applied to chi-square distributed statistics resulting from performing a MH test on multiply imputed data. The differences between these two pooling methods are rather small in terms of statistical power and Type I error rates, and both methods maintain the Type I error rate within a nominal 5% level. When categorical analysis involves tables that are larger than 2×2, the same transformation and pooling procedures can be applied to a (stratified) CMH test for $C_1 \times C_2$ tables. Common odds ratio for 2×2 tables can also be well estimated using the MH method with subsequent log transformation at the pooling stage.

From our simulation study, we have also observed that the statistical power is higher when imputing the underlying continuous endpoint compared to imputing the binary endpoint directly. The former imputation approach also has better power than observed cases analysis under an MAR assumption. Our simulation results suggest that in situations where there is no underlying continuous parameter and where only an odds ratio is of interest for a 2×2 table under MAR, analysis can be performed in an unbiased manner and with smaller power losses using observed cases only. However, even under an MAR assumption, imputations are in general needed to obtain unbiased estimates of other statistics, e.g., proportions and their differences, for which the observed cases analysis could be biased.

It should be noted that in addition to the assumption of normally distributed estimated statistics, Rubin's combination rules also require that the imputation should be proper (i.e., based on a Bayesian posterior predictive distribution), and that the analysis model (applied to multiple imputed datasets) should be compatible with the imputation model used to generate filled-in values (Rubin, 1987). Compatibility can be achieved by using the same likelihood specification in the analysis model as the Bayesian imputation model. For some of the analyses that we discussed in this paper, this is not the case. For example, we performed imputation using an ordinal logistic regression model, but the CMH and MH analyses are not likelihood based. Nevertheless, the existing evidence in the literature indicates that multiple imputation performs well even when this requirement is not satisfied (Shafer, 2003; van Buuren, 2007).

APPENDIX: ADDITIONAL DETAILS ABOUT SIMULATION SETTINGS

Simulated datasets have two treatment arms (P = Placebo and E = experimental) with N=100 subjects per arm.

Longitudinal data was generated with a baseline and 3 post-baseline visits using multivariate normal model for the underlying continuous endpoint with the following correlation structure:

	Y₀	*Y₁*	*Y₂*	*Y₃*
Y₀	1	0.5	0.3	0.2
Y₁	0.5	1	0.5	0.3
Y₂	0.3	0.5	1	0.5
Y₃	0.2	0.3	0.5	1

Derived binary responder status at the final post-baseline visit Z_3 was created based on 50% reduction from baseline, $Z_3=I\{(y_3-y_0)/y_0 < - 0.5\}$.

For the mean structure of the outcome matrix ***Y*** we assumed constant variance over time equal to 1.5, and means at baseline μ_0=10 in both treatment arms. Post-baseline visit means for experimental and placebo subjects were assumed to be linearly reducing from baseline to last visit 3, and calibrated so as to ensure desired probability of binary response $P(Z_3=1|T=\text{"E"})$ and $P(Z_3=1|T=\text{"P"})$ at the last visit according to simulation scenario specifications (see Table 10). To calibrate the means, numerical integration was carried out to solve for μ_3 in the following equation for the specified probability of the binary response:

$$
\begin{aligned}
prob(y_3 < 0.5y_0) &= \int_{-\infty}^{-\infty} f_{y_0}(u)\Phi\left(\frac{0.5u-\mu_{3|0}}{\sigma_{3|0}}\right)du = \\
&= \int_{-\infty}^{-\infty} f_{y_0}(u)\Phi\left(\frac{0.5u-\mu_3-\rho\sigma_3(u-\mu_0)/\sigma_0}{\sigma_3\sqrt{1-\rho^2}}\right)du \qquad (5)\\
&= \int_{-\infty}^{-\infty} \varphi(z)\Phi\left(\frac{z(\sigma_0/2-\rho\sigma_3)+\mu_0/2-\mu_3}{\sigma_3\sqrt{1-\rho^2}}\right)dz
\end{aligned}
$$

Where φ and Φ are standard normal PDF and CDF, respectively; $\mu_{3|0}$ and $\sigma_{3|0}$ are the mean and standard deviation for conditional normal distribution $f(y_3|y_0)$ and ρ is the correlation coefficient between y_3 and y_0. In order to obtain the desired probabilities of response at the last post-baseline visit, the following means were derived from (5):

Probability of response at the last post-baseline visit, *P(Z₃=1)*	Mean at the last post-baseline visit (μ₃)
0.2	6.293607
0.391	5.425333
0.4	5.389407
0.609	4.574672

Under the simulated MAR model of missingness, only subjects whose outcome worsened by some amount equal to or greater than some cut-off value were “eligible” to drop out at post-baseline visits 2 or 3. Once a subject reached this “dropout condition”, s/he was assumed to discontinue with the probability γ. While holding γ=0.5, we calibrated the dropout eligibility cut-off so as to ensure the desired percentage of the dropout in the placebo arm.

	Dropout rate			
Responders rate in placebo arm	10%	20%	30%	40%
	Dropout eligibility cut-off (based on continuous endpoint)			
0.4	9.83	9.049	8.45	7.89
0.2	10.20	9.41	8.84	8.30

REFERENCES

Agresti, A. 2002. *Categorical Data Analysis, Second Edition*. New York: John Wiley & Sons.

Carpenter J.R., Kenward M.G. 2013. *Multiple Imputation and its Application*. Chichester: John Wiley & Sons.

European Medicines Agency. 2010. *“Guideline on Missing Data in Confirmatory Clinical Trials. 2 July 2010. EMA/CPMP/EWP/1776/99 Rev. 1*”. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf

Goria, M.N. 1992. On the forth root transformation of chi-square. *Australian Journal of Statistics*, 34 (1), 55-64.

Landis, R.J., Heyman, E.R., Koch, G.G. 1978. Average Partial Association in Three-way Contingency Tables: A Review and Discussion of Alternative Tests. *International Statistical Review*, 46, 237–254.

Li, K.-H., Meng, X.-L., Raghunathan, T.E., and Rubin, D.B. 1991. Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, 1(1), 65-92.

Mantel, N., Haenszel, W. 1959. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute*, 22, 719–748.

National Research Council. 2010. 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: The National Academic Press.

Rubin D.B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

SAS Institute Inc. 2011. SAS/STAT® 9.3 *User’s Guide.* Cary, NC: SAS Institute Inc..

Shafer, J.L. 2003. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57, 19-35.

Wilson, E.B., Hilferty, M.M. 1931. The distribution of chi-squared. *Proceedings of the National Academy of Sciences*, Washington, 17, 684–688.

Van Buuren, S. 2012. *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC Press.

Van Buuren, S. 2007. Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. *Statistical Methods in Medical Research*, 16, 219–242.

ACKNOWLEDGMENTS

We would like to thank Quintiles Inc. for encouraging and supporting this work as well as conference participation.

We would like to sincerely thank Russ Wolfinger and Yang Yuan for their helpful feedback regarding the application of Wilson-Hilferty transformation.

ERRATA

An initial version of this paper contained errors in equations (3) and (4) and SAS Code Fragment 6 which are now corrected. Simulations have not been re-run.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Bohdana Ratitch
Quintiles Inc.
100 Alexis-Nihon Blvd., Suite 800
Saint-Laurent, Québec, Canada, H4M 2P4
E-mail: bohdana.ratitch@quintiles.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.