

Expert Report on Multiple Imputation*

Donald B. Rubin, Ph.D.

Outline

- 1. Introduction and conclusions**
- 2. Role of missing data in the pivotal trial**
- 3. Use of multiple imputation in the pivotal trial**
- 4. Analysis of various multiple imputation models**
- 5. Missing at random assumption**
- 6. Rate of missing information versus raw quantity of missing data**
- 7. Summary of review**

Introduction and conclusions

This report provides a review of the multiple imputation methodology used by the Sponsor (Acorn Cardiovascular, Inc.) to handle missing data in the analysis of the primary endpoint of its pivotal clinical trial of the CorCap Cardiac Support Device (CorCap), as well as an evaluation of the impact of multiple imputation on the results of that analysis. This included a completely blinded imputation briefly described at the end.

The present review was prompted by concerns raised to the Sponsor by the Food and Drug Administration's (FDA) Office of Device Evaluation (ODE), as well as by proceedings at the FDA Advisory Panel convened in June 2005 to consider the Sponsor's Premarket Approval Application (PMA 040049) for the CorCap. In particular, sources of relevant information to which this report is directed are the not approvable letter issued by ODE to the Sponsor on August 12, 2005, ODE's formal review of the statistical methods used in the Sponsor's PMA (FDA Memorandum of Review v5.2, May 23, 2005), and comments made during the course of the Advisory Panel meeting held June 22, 2005.

Specifically, this report addresses ODE's concern that the use of multiple imputation may have "compromise[d] the analysis of the primary endpoint" (FDA's questions to Advisory Panel, Panel transcript p. 80), potentially by permitting the Sponsor to make an inappropriate claim of statistical significance comparing treatment to control. The present report shows that, contrary to that reviewer's conjecture, several different analyses of the trial outcomes result in essentially the same conclusion, and that the results are not compromised by the use of multiple imputation to handle missing data. Documents and details of previous analyses were provided to Donald B. Rubin, Ph.D. (DBR) by Scott Brown, Ph.D. (SB), and the new SAS computations implemented by SB. Moreover, an entirely new blinded multiple imputation analysis was designed by DBR, but implemented by Constantine Frangakis, Ph.D. (CF) who was blinded to the study, the treatments, etc. The method used was a state-of-the-art method developed for the Centers for Disease Control and Prevention (CDC) for use in their anthrax vaccine randomized trials to be presented to FDA after those trials are completed; DBR and CF worked together with others to develop and implement this approach for CDC. The results based on this new analysis were entirely consistent with the previous ones and were even more significant.

* Scott Brown, Ph.D., statistician for Acorn Cardiovascular, Inc., provided all details on early imputation efforts.

Role of missing data in the pivotal trial

The appropriate statistical handling of missing data requires an understanding of its source and structure. In the Sponsor's pivotal trial, most of the missing data were introduced by the timing of an agreement with ODE that strongly encouraged, after trial enrollment had begun, the use of a blinded instrument to measure New York Heart Association (NYHA) classification of cardiac functional status.

This agreement was motivated by ODE's concern that use of the standard NYHA classification (that is, an instrument assessed by the attending physician) could result in bias because the attending physician could not be blinded due to ethical and other practical reasons.

The Sponsor's primary endpoint was a clinical composite of all-cause mortality, incidence of major cardiac procedures (MCPs) – a list including mitral valve repair or replacement, tricuspid valve repair or replacement, bypass graft surgery, implantation of a left ventricular assist device, and cardiac transplant – indicative of worsening heart failure during study follow-up, and change from baseline to final follow-up visit in blinded NYHA classification. This primary endpoint was missing baseline data for blinded NYHA classification in the 174 out of 300 patients (58%) who enrolled before the agreement was implemented.

Multiple imputation was recommended by ODE (FDA Response to Acorn IDE G990267 Supplement #44, dated May 19, 2004) as an acceptable method of handling the resulting missingness. Specifically, ODE's response cited its position that “[f]or primary analyses, it is not acceptable to use two unblinded assessments or to use one blinded and one unblinded assessment. We recognize that using the two blinded assessments entails using an incomplete data set. Therefore, we recommend statistical imputation of the missing baseline data prior to computing the change score for these patients. ... We recommend that multiple imputations be done as opposed to a single imputation in order to assess variability in the imputations.”

Use of multiple imputation in the pivotal trial

Adhering to the intention-to-treat principle in such situations requires some form of imputation, either explicit or implicit. Multiple imputation is a technique for handling the statistical analysis of data with missingness, and is widely regarded as a superior alternative to simpler methods such as casewise deletion, replacement by mean values, weighting adjustments, single imputation and other methods (Little and Rubin, 2002, chapters 1, 3 and 4). These other methods do not, in general, preserve valid statistical inference, because estimates of treatment effect may be biased and confidence intervals almost always inappropriately narrow. Among the consequences of using such methods are anticonservative p-values, with inappropriate declarations of statistical significance.

Multiple imputation, on the other hand, preserves valid statistical inference by simulating several different sets of imputed data and estimating both the variance within each imputation (i.e., as if the imputations were exactly correct) and the variance between imputations (i.e., reflecting the uncertainty of the imputations). The combination of answers from all imputed datasets results in a valid (approximately unbiased) estimate of the true treatment effect, while simultaneously producing valid confidence intervals with appropriate coverage and accurate p-values. This conclusion is supported both by theoretical evaluations and extensive simulation results (e.g., see references in Rubin, 1996).

In the Sponsor's pivotal trial, multiple imputation was implemented using a model developed with the input of outside statistical experts not including DBR, and subsequently validated by additional imputation models developed at the behest of ODE's statistical reviewer, as well as models developed by DBR, including the blinded models implemented by CF.

When applied to the pivotal trial data, all multiple imputation models render the same conclusion: that patients randomized to CorCap had significantly better clinical outcomes, as defined by the primary endpoint, than patients randomized to control, meeting an *a priori* success criterion. In fact, the blinded imputation resulted in the most significant p-value for the primary endpoint.

Analysis of various multiple imputation models

Original analysis from Sponsor's PMA

In the original analysis presented in the Sponsor's PMA, baseline values for the blinded core lab NYHA instrument were multiply imputed as follows:

Model 0

- a. A group of 10 potential explanatory variables collected at baseline (pre-randomization) were entered into a stepwise regression, ultimately producing a set of four baseline predictors of blinded (core lab) NYHA.
- b. The four predictors – site NYHA, MLHF score, 6-minute walk distance and SF-36 physical functioning score at baseline – were employed in a multiple imputation model (SAS PROC MI) of core lab NYHA at baseline, using Markov chain Monte Carlo simulation to produce 100 imputed datasets. This model assumes a normally-distributed dependent variable, and because the NYHA instrument is actually an ordinal response with four levels, the resulting imputed values were rounded off to the nearest whole number for the purpose of further analysis, a procedure recommended, for example, by Schafer (1997).
- c. One hundred imputed datasets were generated and collectively analyzed (SAS PROC MIANALYZE) to produce parameter estimates and p-values for the primary endpoint.

Model 0 yielded a primary endpoint demonstrating a significant difference between treatment and control ($p=0.024$), strongly supporting superiority of treatment over control.

Analyses conducted during the PMA review process

The Sponsor's PMA model was reviewed by ODE, which expressed the following concerns:

- The normal-distribution assumption was unnecessary, because it is feasible to model NYHA classification as an ordinal variable;
- The set of predictor variables chosen was too narrow – in particular, it did not include several variables that were ultimately used in the analysis of the primary endpoint,

meaning that the “imputer’s model” did not encompass the “analyst’s model,” violating one of the guidelines recommended in Rubin (1987; 1996; 2004);

- The use of stepwise regression to narrow the set of predictors produced a source of variability in the modeling that was unaccounted for in the Sponsor’s analysis.

To address ODE’s concerns, initially two imputation models were developed using different assumptions.

Model 1

- a. All 10 of the original PMA predictors, as well as a set of variables requested by the ODE statistical reviewer, were included in this imputation model. The list of variables is: baseline site NYHA, baseline 6-minute walk, baseline MLHF score, baseline SF-36 PF score, age at randomization, gender, ischemic etiology, time from heart failure diagnosis to randomization, LVEF at baseline, baseline peak VO₂, baseline DBP, MVR vs. no-MVR stratum, size of clinical site (large/medium/small), 12-month site NYHA, death, and presence of major cardiac procedures during follow-up. Note that unlike the PMA model, this modified imputation model included data collected after randomization, which is valid and usually results in more precise estimates.
- b. The variables above were used to multiply impute core lab NYHA at baseline, using a logistic regression model in which the response was considered an ordinal, categorical variable. Predictor variables themselves with missing values were initially imputed using an MCMC model in SAS, identical in structure to the one employed in Model 0, including all variables except for core lab NYHA, in order to eliminate missing values before imputing core lab NYHA.
- c. As before, the 100 resulting imputed datasets were analyzed (SAS PROC MIANALYZE) to produce parameter estimates and p-values for the primary endpoint.

When applied to the Sponsor’s primary endpoint, this imputation model also resulted in a significant difference between treatment and control ($p=0.029$).

Model 2

Also prior to DBR’s involvement, a second modified imputation model was built to address further FDA concerns. Specifically, this model again used an ordinal assumption on the outcome variable of baseline core lab NYHA, and it included the following 14 predictors: length of follow-up, gender, peak VO₂ at baseline, diastolic blood pressure at baseline, MVR stratum, size of clinical site (small/medium/large), duration of heart failure at baseline, age at randomization, baseline 6-minute walk distance, baseline MLHF score, baseline SF-36 physical functioning score, ischemic etiology at baseline, LVEF at baseline, and baseline site-assessed NYHA class. This model, Model 2, is therefore similar to the preceding model (Model 1), except that it contains no data collected after randomization, and was again implemented in SAS to create 100 multiply imputed datasets, which were then analyzed using SAS.

The results of this second modified analysis again showed significance when comparing treatment to control on the primary endpoint ($p=0.033$).

Analyses conducted for the present review

Model 3

For the purposes of this report involving DBR, another imputation model was constructed using a similar set of predictors as the preceding two models. Specifically, the predictors included gender, peak VO₂ at baseline, diastolic blood pressure at baseline, MVR stratum, size of clinical site (small/medium/large), duration of heart failure, age, baseline 6-minute walk distance, baseline MLHF score, baseline SF-36 physical functioning score, ischemic etiology, LVEF at baseline, baseline site-assessed NYHA class, and site-assessed NYHA class at 12 months. This list is identical to the set of variables in Model 1, except that length of follow-up was omitted from consideration due to difficulties with convergence in simulation of the MCMC chains for creating multiple imputations.

As in the preceding models, the core lab NYHA instrument was treated as an ordinal response in SAS. The key difference between this model and the preceding ones was that two separate imputation models were developed, one for each randomized treatment group. This provision eliminated the possibility of “cross-contamination” of imputation models between randomized arms, thereby permitting separate modeled relationships between the core lab NYHA classification and the various predictors by treatment group.

The results of this imputation model agreed with the preceding ones; in particular, the difference between treatment and control in the primary endpoint was found to be positive and significant ($p=0.021$).

Model 4

Lastly, a final imputation model was designed by DBR and implemented in a blinded fashion by a third party (CF) who did not have access to identifying information in the data – including which group was treatment and which was control – or knowledge of the identity of the sponsor company. The dataset provided for this version of the imputation was also denatured to eliminate informative names on predictor and response variables, to further assure a blinded evaluation of the data. Over 100 variables at both baseline and follow-up were provided for this version of the multiple imputation.

Like Model 3, this imputation procedure was conducted separately for each randomized group to avoid “cross-contamination,” and using the same modeling principles in each group. These principles included the following:

- Data were regarded as “observed,” “missing,” or “undefined,” the latter if the scheduled measurement occurred after either death or after a major cardiac procedure indicative of worsening heart failure. Only missing data, not undefined data, were imputed.
- The time that would determine if the data would be missing or undefined – that is, the earliest of either death or major cardiac procedure – is referred to as the “failure time.” The failure time is missing if a subject entered later than the earliest entry time and had not died or experienced a major cardiac procedure before the study’s common closing date. It was assumed that patients who entered the study at different times were, conditional on modeled covariates, exchangeable in their inherent characteristics, which implies that missing information of the failure time is missing ignorably. Missing failure

times were imputed from a Bayesian predictive distribution obtained from the Kaplan-Meier likelihood.

- The imputation assumed that remaining missing data were missing ignorably based on the other observed data and based on the completed (observed and imputed) failure times. With this principle, any data that were actually observed but were deemed “undefined” because they occurred after a major cardiac procedure were not used as part of the observed data for the imputation.

The blinded imputation results were returned as a set of five imputations of core-lab NYHA. The result of incorporating this set of imputations into the primary endpoint analysis concurred with all preceding results, in that superiority of treatment to control was demonstrated ($p=0.014$).

Missing at random assumption

All of the multiple imputation models above make the assumption that data were missing at random (MAR). The MAR assumption cannot be directly tested statistically, and is typically justified by examining the circumstances under which missing data arose. Data not MAR are termed “nonignorable,” and require modeling assumptions specific to the distribution of the probability of missingness as a function of values that are missing given the values that are observed.

In the present case, there is no reason to doubt the MAR assumption. Missing data arose in the Sponsor’s pivotal trial, not because of not understood patient dropout or other potential relationships with outcomes, but simply because the blinded NYHA instrument had not been implemented until a number of patients had already been enrolled. In a randomized trial, this entails very strong structural evidence supporting the MAR assumption, and no analysis performed by either the Sponsor or FDA, or argument presented by either, has produced any reason for concern.

Additionally, the fact that the missing data occurred at baseline – prior to randomization – and not during study follow-up allays concerns that the missingness could be related to patients’ clinical courses during the trial. Even if early- and late-enrolled patients differed systematically in some way related to outcomes, the randomization was blocked in time, which prevents any untoward effect of the missing data on any analysis estimating treatment effectiveness relative to control.

Rate of missing information versus raw quantity of missing data

The claim that the Sponsor’s missing data were MAR conflicts with informal opinions expressed by the Advisory Panel statistician during Panel proceedings that the Sponsor’s missing data were, instead, nonignorable. Those statements appeared to be based primarily on an assertion that the quantity of missing data in one baseline component of the primary endpoint (i.e., core lab NYHA) rendered the missingness nonignorable as a matter of fact.

This assertion is incorrect. First, the MAR assumption is not related to the quantity of data missing. MAR is a structural condition dependent upon the statistical relationship between the missingness indicator and missing data conditionally given the observed data, not upon the amount of data that may be missing. No matter what the raw fraction of missing data, this

fraction is not a basis for questioning the MAR assumption. To assert otherwise, is to assert, for example, that a small simple random sample drawn from a large population would create nonignorable missing data simply because most of the population is missing (i.e., because it is not part of the sample that is drawn).

Second, statements made during the Panel's proceedings that over 50% of data were missing in the primary endpoint are factually incorrect. Although it is true that over 50% of data were missing at baseline for the blinded NYHA instrument, these data were only one element in a change score for NYHA, which was in turn one of three components of the Sponsor's composite primary endpoint. In fact, due to the structure of the primary endpoint, which classified patients as "worsened" if death or an MCP occurred regardless of the change in the blinded NYHA instrument, only 104 (35%) patients required an imputed NYHA score at baseline to compute the primary endpoint. Remaining patients without baseline core lab NYHA would have been classified as "worsened" regardless of their change in NYHA, due to the occurrence of death or MCP.

Furthermore, even the figure of 35% substantially overestimates the impact of missingness on the analysis of the primary endpoint. When missing values can be accurately predicted from nonmissing data, the impact of missingness is minimized by the use of valid statistical modeling. The appropriate statistical measure of the impact of missing data is the *fraction of missing information* about the quantity being estimated, which is given by the variability between imputations – i.e., the amount of variability contributed by uncertainty in the imputation of the missing values – as a fraction of the total variability in the analysis. The smaller this ratio, the less impact the missing data have on the results of the analysis.

In the present case, the fraction of missing information was found to be 0.09 when examining the Sponsor's model in the original PMA, a low figure that supports the relative unimportance of the missing baseline NYHA data, if handled correctly, as by appropriate multiple imputation. Subsequent multiple imputation models performed on the same dataset showed fractions of missing information between 0.05 and 0.16, all relatively small values.

As an illustration of the difference between the quantities "the fraction of missing data" and "the fraction of missing information," consider a dataset in which height in inches (rounded to the nearest inch) is desired, but that half of those values are randomly missing. Now suppose that in the same dataset, height in centimeters (rounded to the nearest centimeter) is always present. In this case, the missing data can be predicted with great accuracy from the observed data, and although the fraction of missing data is 50%, the fraction of missing information is near zero for any quantity involving height in inches (e.g., the 25th percentile of height).

The conclusion is that the specific structure of the multiple imputation process was very unlikely to have influenced the statistical analysis of the primary endpoint. This is opposite to the critical opinion expressed during the Advisory Panel proceedings that, due to the quantity of missing data present, the specific statistical model used would drive the results. Understanding the difference between the fraction of missing data and the fraction of missing information provides the logical reason for the substantial agreement found among the various imputation methods.

Summary of review

The validity of the Sponsor's trial results were questioned by both ODE and the Advisory Panel convened in June 2005 to consider the use of multiple imputation to handle missing data. Specifically, concerns were raised by ODE that the use of multiple imputation may have "compromise[d] the analysis of the primary endpoint," while opinions expressed during the Advisory Panel proceedings also cast doubt on the soundness of the Sponsor's analyses. Criticisms of the Sponsor's PMA results rested on several bases:

1. That the quantity of missing data in the primary endpoint was great – over 50 percent – and this rendered the analysis of the primary endpoint highly susceptible to the specific imputation model employed, limiting confidence in the PMA results;
2. That the quantity of missing data alone provided a factual basis for claiming nonignorability of missingness, and that this therefore rendered the Sponsor's imputation results invalid, because they relied upon the MAR assumption;
3. That the Sponsor's PMA imputation model was insufficiently broad in terms of the predictors included, and that it made unnecessary distributional assumptions regarding the response variable, which could have resulted in inaccurate results.

With regard to criticism number 1, the quantity of missing data in the primary endpoint has been shown to be substantially less than was claimed during the Advisory Panel, and that furthermore, an appropriate statistical evaluation of the impact of that missingness has shown that the true effect of the missing NYHA on the primary endpoint was quite small.

With regard to criticism number 2, the quantity of missing data alone cannot be a basis for claiming nonignorability, and in this particular case, there is no reason to doubt the MAR assumption.

With regard to criticism number 3, although the Sponsor's PMA model may not have been ideal for the multiple imputation of the missing data, various other multiple imputation models with appropriate distributional assumptions and sufficiently broad sets of predictor variables, all support the conclusion drawn in the PMA: that patients randomized to CorCap had significantly better clinical outcomes, as defined by the primary endpoint, than patients randomized to control.

References

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data*. J. Wiley & Sons, New York. (2002) Second edition.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons. (2004) Classic edition.

Rubin, D.B. (1996) Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91, 473-489.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.