

# *Columbia University*

Columbia University Biostatistics Technical Report Series

---

*Year 2000*

*Paper 16*

---

## Losses To Follow-Up In WARSS Collaboration Datasets: A Detailed Statistical Presentation of the Imputation Procedures

John L.P. Thompson\*

Bruce Levin†

\*

†

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/columbiabiostat/art16>

Copyright ©2000 by the authors.

# Losses To Follow-Up In WARSS Collaboration Datasets: A Detailed Statistical Presentation of the Imputation Procedures

John L.P. Thompson and Bruce Levin

## **Abstract**

This document prospectively records the procedures which will be used for handling losses to follow-up (LTF) in statistical analyses of WARSS data. They were developed by B Levin Ph.D. (WARSS senior statistical consultant) and JLP Thompson Ph.D. (WARSS statistician), and have been approved by the SOCC (Statistical Oversight and Coordinating Committee of the WARSS collaboration). They have been accepted by the WARSS Principal Investigator (JP Mohr M.D.), and also by the Principal Investigators of APASS, PICSS, HAS, and GENESIS for use in these collaborating studies.

# Losses To Follow-Up In WARSS Collaboration Datasets: A Detailed Statistical Presentation of the Imputation Procedures

J. L. P. Thompson Ph.D., B. Levin Ph.D.  
Revised 7/28/00

## A. INTRODUCTION

This document prospectively records the procedures which will be used for handling losses to follow-up (LTF) in statistical analyses of WARSS data. They were developed by B Levin Ph.D. (WARSS senior statistical consultant) and JLP Thompson Ph.D. (WARSS statistician), and have been approved by the SOCC (Statistical Oversight and Coordinating Committee of the WARSS collaboration). They have been accepted by the WARSS Principal Investigator (JP Mohr M.D.), and also by the Principal Investigators of APASS, PICSS, HAS, and GENESIS for use in these collaborating studies.

Section B summarizes the proposed approach. Section C gives the reasoning behind it, and the reasons for preferring it to some possible alternatives. Section D presents some of the procedures in more detail.

## B. SUMMARY

1. Using relevant clinical data and data on putative reasons for WARSS LTFs, classify them into 3 groups:

I) ***Endpoint is “imminent.”*** There is evidence at LTF of high risk of primary endpoint (e.g., dramatically worsening symptoms).

II) ***LTF is not independent of endpoint.*** Endpoint is not imminent, but evidence suggests that the cause of the LTF is not statistically independent of time to future event.

III) ***LTF is independent of endpoint.*** Endpoint is not imminent, but evidence suggests that the cause of the LTF is statistically independent of time to future event.

Where there is uncertainty whether an LTF should be coded I) or II), code it I).

Where there is uncertainty whether an LTF should be coded II) or III), code it II).

Groups I) and II) represent determinations that the remaining follow-up time for an LTF is non-ignorably missing, i.e., that the mechanism causing the loss depends on the missing data.

Group III) represents a determination that the remaining follow-up time for an LTF is ignorably missing, i.e., that the mechanism causing the loss is statistically independent of the missing data.

- 2 Complete the dataset by handling the three LTF groups as follows:
  - I) *Endpoint is imminent*: Impute the occurrence of a primary endpoint at the point of LTF.
  - II) *LTF is not independent of endpoint*: Develop a clinically reasonable model which imputes missing events and/or follow-up time for these cases. The time-to-primary-endpoint model utilizes a parameter for elevated relative risk compared to other subjects with complete data, reflecting the non-ignorable missingness. The model also incorporates baseline factors that affect risk.
  - III) *LTF is independent of endpoint*: Censor these cases at the point of LTF.
- 3 Analyze by repeated imputation. A completed dataset comprises the data from the cases which are not LTF (i.e., follow-up is complete) combined with the imputed and censored data for the three LTF groups. For any such completed dataset, calculate all primary and secondary statistics such as P-values, point estimates, confidence intervals, etc. In order to reflect uncertainty in what the missing data “might have been” had they been observed, the entire procedure is repeated 25 times allowing different imputations for group II) LTFs according to the time-to-primary-endpoint model. Methods for combining the repeated imputations are presented below. For the primary and secondary statistical analyses for WARSS and the collaborative studies, and all other WARSS analyses of follow-up, reports will summarize the results of repeated imputations from the model.
- 4 Sensitivity analyses assess how the WARSS primary result changes over a range of clinically convincing alternative models for the missing data. Similar sensitivity analyses are also conducted for the primary analyses for the collaborative trials.

## C. REASONING

Overall, we adopt a composite approach that distinguishes the different types of LTF and incorporates clinical and statistical assumptions that are appropriate for each type. The intent-to-treat principle is preserved throughout.

### *Non-ignorably missing data*

Many approaches are possible for imputing non-ignorably missing data. Each has particular strengths and weaknesses. For our purposes it is useful to distinguish three.

1. State the upper and lower limits.

The upper limit uses “worst/best” imputation. In WARSS, this involves imputing an immediate event at LTF for one treatment group, and event-free follow-up to two years for the other treatment group. The lower limit uses “best/worst” imputation, which reverses the above imputations for the two groups.

The advantage of this approach is that it provides solid upper and lower limits between which the “real” result (i.e., the one that would have occurred had there been no LTFs) must fall. This is usually suitable if missing data are very limited, or if they are less limited but accompanied by a very large treatment effect (e.g., if there are 20 LTFs but a difference of 100 events between the two treatment groups).

The major drawback is that if the missing data are not minimal, the limits encompass a very wide range of results. In addition, some of the results that fall within the limits reflect assumptions that are clinically unrealistic.

## 2. Impute the worst outcome to both groups.

A primary endpoint is assigned at the moment of LTF, irrespective of treatment assignment. This “worst/worst” imputation produces just one result for the worst possible outcome. It has the greatest impact on the group with more losses to follow-up (if there is one). It is reasonable where missingness is highly indicative of impending endpoint. When missing data are very limited, it does not have a great impact, but can be reassuring. An example is the NINDS tPA trial. It may not be clinically plausible there, but this does not really matter because the number of LTFs was so small.

The possible disadvantages emerge where the amount of missing data is non-trivial. If LTFs are not for reasons related to illness, penalizing the treatment arm with the most LTFs is potentially misleading and therefore inappropriate. In addition, the approach may not be clinically convincing if the group with less expected adverse events has more LTFs.

## 3. Develop a model to impute the missing data.

The goal is to develop a model for imputing missing events and/or follow-up time for LTFs which are clinically reasonable, and thus will yield a clinically acceptable point estimate of the relative efficacy of  $T_1$  vs.  $T_2$ . This approach also facilitates clinically reasonable sensitivity analyses that allow for an appropriate range of assumptions. The hope is to be able to say that for the results of the analysis to be wrong, one would have to make assumptions for LTFs that are clearly dismissable to the overwhelming weight of clinical opinion. That statement would be consistent with the ethical goal of a clinical trial to perturb the state of clinical equipoise.

The disadvantages are that the result is only as valid as the model which produces it, and it is not possible to validate the model directly from the data collected in the trial in question. But if sensitivity analyses show that for the model to be wrong, highly questionable clinical assumptions would have to be made, the approach gains credibility.

In WARSS, there are too many non-ignorably missing data to adopt either approach 1 or approach 2 globally. Each would yield a conclusion that is too massively affected by clearly unrealistic clinical assumptions. We therefore recommend a combined approach, imputing

immediate events (approach 2) for group I) and model-based imputation (approach 3) for group II). There is substantial knowledge about the nature of the effects of warfarin and aspirin, and this will facilitate the development of a credible model. Sensitivity analyses will help to assess whether the trial results are clinically reasonable.

### *Ignorably missing data*

Every imputation of missing data requires some assumption(s). Data are ignorably missing when the cause of missingness can be assumed to be statistically independent of the missing data, e.g., the future time of primary endpoint. When this single strong assumption holds, the case can be treated as a censored observation.

Censoring data that are ignorably missing has major advantages. It uses all of the legitimate follow-up information collected in the trial; it does not require any assumptions or imputations regarding future follow-up; and it is valid statistically because it does not introduce bias.

We adopt the censoring approach for Group III) WARSS LTFs. Since it is crucial that the independence assumption be tenable, we are conservative in making it.

## **D. THE PROCEDURES IN MORE DETAIL**

Since the procedures for groups I) and III) are straightforward, they are not discussed here. The other key steps are:

- A. Specification of the imputation model for group II)
- B. Statistical analyses that combine the observed and the imputed data
- C. Sensitivity analyses

### A. Specification of the imputation model for group II).

Two techniques can be used to impute missing data, the first under the null hypothesis of no treatment effect, and the second under the alternative hypothesis that there is a difference between warfarin and aspirin. We begin by specifying the first imputation technique to be used for the primary test of the null hypothesis. If and only if there is a significant difference (see section B for the significance test in the presence of missing data), the missing data will be imputed again under the second technique for best estimates of treatment effects. The first two steps of the procedure are the same under either technique.

- i) The project clinicians identify the baseline (pre-randomization) risk factors likely to be related to the probability of an outcome event.
- ii) Using existing trial results (e.g., Clinical Trialists' Collaboration on Aspirin), specify a parameter,  $\gamma$ , that represents the log relative hazard odds ratio for group II) LTFs compared to those who are not LTF. We propose to use the hazard odds ratio  $e^\gamma = 1.2$  for LTF vs not LTF.

Imputation under the null hypothesis:

- iii) Using the actual WARSS subjects with complete follow-up (primary endpoints included), draw a bootstrap sample, pooling together warfarin and aspirin subjects. Suppose there are  $n$  subjects in total with complete follow-up. A bootstrap sample consists of drawing  $n$  subjects at random and *with* replacement from among all those patients with complete follow-up.
- iv) With the bootstrap sample, fit a discrete-time Cox model of the form

$$\frac{h(t|x)}{1-h(t|x)} = e^{\beta x} \frac{h(t|0)}{1-h(t|0)},$$

where time  $t$  is measured in days of follow-up, and  $h(t|x) = P[T = t | T \geq t, x]$  is the discrete-time hazard probability that failure occurs on day  $t$  of follow-up given no prior failure, for a subject with a vector of covariates  $x$ . The reference group has  $x=0$ . After maximum partial likelihood estimation of  $\beta$ , estimates of the survival function  $S(t|x)$  for subjects with covariate vector  $x$  are obtained using maximum likelihood logistic regression to estimate the reference log hazard odds  $h(t|0)$  at each time  $t$  of an observed primary endpoint.

- v) The next step is to obtain the future or residual follow-up time distribution for WARSS patients not LTF after any given observed duration of follow-up. For a subject with baseline covariate vector  $x$ , the residual waiting time survival function beyond observed follow-up time  $t_0$  is estimated by

$$\hat{S}(t|t \geq t_0, x) = \prod_{t_0 < i \leq t} \left( 1 + e^{\hat{\beta} x} \frac{\hat{h}(i|0)}{1 - \hat{h}(i|0)} \right)^{-1}.$$

- vi) To reflect the non-ignorable missingness of the WARSS patients who are LTF, we increase the exponent from  $\hat{\beta} x$  to  $\gamma + \hat{\beta} x$  in the above expression.
- vii) Generate a random variable  $t^*$  using  $\hat{S}(t|x, t_0)$  for each subject LTF given his or her specific covariate  $x$  and observed follow-up  $t_0$ , and use this to impute the missing follow-up data. If  $t_0 + t^* \geq 2$  years (i.e., 761 days), the subject is imputed as complete without event. Otherwise the subject is imputed as having had a primary endpoint at time  $t_0 + t^*$ .

- viii) The imputed data are added to the observed WARSS data to constitute one completed data set. The entire process, steps iii) through vii), is then repeated to produce 25 completed data sets.

Imputation under the alternative hypothesis:

- iii) Using the actual WARSS subjects with complete follow-up (primary endpoints included), draw a bootstrap sample, separately for warfarin and aspirin patients. Suppose there are  $n_1$  warfarin subjects and  $n_2$  aspirin subjects with complete follow-up. A bootstrap sample consists of drawing  $n_1$  subjects at random and *with* replacement from among all those warfarin patients with complete follow-up, and similarly, drawing  $n_2$  subjects at random and with replacement from among all those aspirin patients with complete follow-up.
- iv) With the bootstrap samples, fit a discrete-time Cox model of the form

$$\frac{h_j(t|x)}{1-h_j(t|x)} = e^{\beta_j x} \frac{h_j(t|0)}{1-h_j(t|0)},$$

where  $j=1,2$  indicates treatment arm, time  $t$  is measured in days of follow-up, and  $h_j(t|x) = P[T = t | T \geq t, x, j]$  is the discrete-time hazard probability that failure occurs on day  $t$  of follow-up given no prior failure, for a subject in treatment group  $j$  with a vector of covariates  $x$ . The reference group has  $x=0$ . After maximum partial likelihood estimation of  $\beta_j$  for each treatment arm, estimates of the survival functions  $S_j(t|x)$  for subjects with covariate vector  $x$  are obtained using maximum likelihood logistic regression to estimate the reference log hazard odds  $h_j(t|0)$  at each time  $t$  of an observed primary endpoint.

- v) The next step is to obtain the future or residual follow-up time distribution for WARSS patients not LTF after any given observed duration of follow-up. For a subject with baseline covariate vector  $x$  assigned to treatment group  $j$ , the residual waiting time survival function beyond observed follow-up time  $t_0$  is estimated by

$$\hat{S}_j(t|t \geq t_0, x) = \prod_{t_0 < i \leq t} \left( 1 + e^{\hat{\beta}_j x} \frac{\hat{h}_j(i|0)}{1 - \hat{h}_j(i|0)} \right)^{-1}.$$

- vi) To reflect the non-ignorable missingness of the WARSS patients who are LTF, we increase the exponent from  $\hat{\beta}_j x$  to  $\gamma + \hat{\beta}_j x$  in the above expression.
- vii) Generate a random variable  $t^*$  using  $\hat{S}_j(t|x, t_0)$  for each subject LTF given his or her specific covariate  $x$  and observed follow-up  $t_0$ , and use this to impute the missing follow-up data. If  $t_0 + t^* \geq 2$  years (i.e., 761 days), the subject is imputed as complete without event. Otherwise the subject is imputed as having had a primary endpoint at

time  $t_o+t^*$ .

- viii) The imputed data are added to the observed WARSS data to constitute one completed data set. The entire process, steps iii) through vii), is then repeated to produce 25 completed data sets.

## B. Statistical Analyses.

Since  $t^*$  is a random variable, different values will occur in the repetitions of step vii). For *descriptive purposes*, we will present the median of the 25 survival curves for warfarin and aspirin patients. We will also display various quantities of interest, e.g., the log-rank statistic P-values and the log relative hazard parameters comparing warfarin vs. aspirin for the 25 individual completed datasets. Which imputation technique is used for these descriptive purposes depends on the outcome of the primary hypothesis test.

For the *primary hypothesis test*, a special permutation analysis is available thanks to the randomized design of WARSS. In brief, with no missing data, one computes the permutation distribution of the numerator of the log-rank z-score (number of observed minus expected primary endpoints), obtained by considering all possible equally likely random allocations of treatments to subjects. Then the P-value is the two-sided tail-probability in this distribution of the log-rank numerator corresponding to the actual allocation. With missing data, for each of the theoretical allocations, we use the mean of the 25 log-rank numerators from the corresponding completed datasets using the first imputation technique. **This will produce the primary WARSS P-value result.**

We note that there is a distinction between the primary WARSS P-value so obtained and the mean of the log-rank P-values obtained from the 25 individual completed datasets. The latter is an estimate of what the P-value would have been if the data had been complete, but that is not the same as a valid P-value to be used for inference in the presence of missing data.

If there is a statistically significant result in the primary hypothesis test, we will re-impute under the second imputation technique. We will then use Rubin's multiple imputation method to *estimate* key quantities of interest, e.g., the log rate ratio (log hazard odds ratio) for warfarin vs. aspirin, with appropriate standard errors and confidence intervals. These will be reported together with the redone descriptive displays described at the beginning of this section.

## C. Sensitivity Analyses.

We will see how the primary result changes over a range of clinically convincing models.