

# Department Seminars

## Impute Or Not: When Data Are Missing Completely At Random?



**Po-Huang Chyou PhD**

Biomedical Informatics Research Center  
Marshfield Clinic Research Foundation

### Abstract:

**Background:** It is debatable whether imputation is required for missing data. Some researchers suggested complete-subject analysis (no imputation) is appropriate if  $\leq 30\%$  data are missing. Others reported problems of bias with some imputation methods. These arguments are used against imputation. This study investigates if imputation is necessary and if proposed methods give better results.

**Methods:** Existing methods include no imputation, mean, and multiple imputations. Proposed methods include ZP1-2, ZP1-2-sqr, ZP5-6, and ZP5-6-sqr derived by 3-step weighted means from the observed data and random selection. The simulation study is carried out by SAS PLAN procedure.

**Results:** Our data showed that no and mean imputations significantly over- or under-estimated the confidence width of the estimated mean or relative risk, given 10%-90% of the data missing. The proposed methods provide a better confidence width. From simulations with 10-30% missing, no and mean imputations have a  $< 60\%$  accuracy rate on the confidence interval and width measurements, the multiple imputation and ZP5-6 increased it to 67% and 89%, respectively.

**Conclusions:** Imputation improves accuracy and our methods performed better than the popular multiple imputation method under the assumption of missing completely at random. Further theoretical work may enhance the proposed methods.

Key Words: Imputation, Missing completely at random, Delete, Mean, Confidence width, Simulation.

**Friday,**

**October 9, 2009**

**G5/113 CSC**

**12:00-1:00 p.m.**

Seminars sponsored by the Department of



Biostatistics & Medical Informatics