

# Performing Sensitivity Analyses of Imputed Missing Values

Huifang Qin  
Mike Singleton

Kentucky CODES  
Kentucky Injury Prevention & Research Center  
University of Kentucky

July 14<sup>th</sup>, 2003

[www.kiprc.uky.edu](http://www.kiprc.uky.edu)

29<sup>th</sup> TRF 2003, Denver

**Multiple Imputation in Public Health Research**

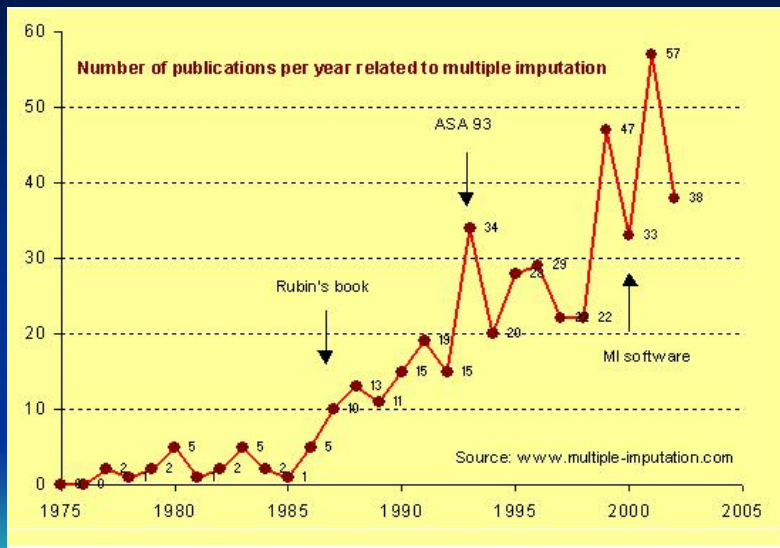
**Handling Missing Data in Nursing Research with Multiple Imputation**

**NHTSA: Transitioning to Multiple Imputation**

**A new Method to Impute Missing BAC values in FARS**

**Application of Multiple Imputation in Medical Studies: from AIDS to NHANES**

**Multiple Imputation Publications**



July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## What's Multiple Imputation

Multiple imputation is a statistical technique for analyzing incomplete data sets, that is, data sets for which some entries are missing. Application of the technique requires three steps: imputation, analysis and pooling.

**Incomplete data sets** --- data set with missing values (>5%)

**Imputation** --- 'filling in' missing data with plausible values and create *n* imputed data sets

**Analysis** --- Analyze every imputed data set

**Pooling** --- Integrate the analysis results into a final result.

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## Questions???

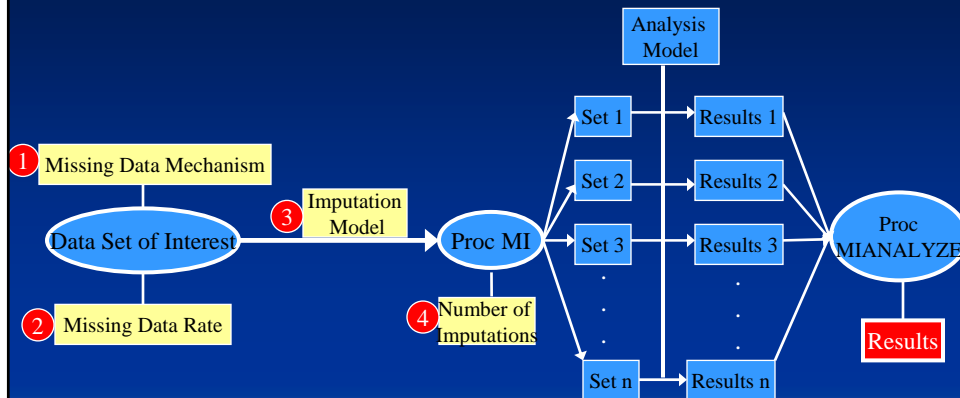
- May I use MI to deal with missing data problems for my data sets?
- How can I believe that the MI will give me better analysis results?
- What should I do to get good results from MI?

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## MI Process and Factors that Affect the Results



July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## CODES Application

### Research Question:

What was the relationship between driving under the influence of drugs and/or alcohol (DUI), and being killed or hospitalized (K/H) in a crash, for motorcycle riders in Kentucky in 2001?

### Outcome (Dependent Variable):

Killed or Hospitalized (K/H)

### Risk Factor Candidates (Independent Variables):

Age, gender, suspected DUI, posted speed limit, helmet use,  
fixed object, head-on collision, collision time, rural vs. urban

July 14<sup>th</sup>, 2003

[www.kiprc.uky.edu](http://www.kiprc.uky.edu)

29<sup>th</sup> TRF 2003, Denver

## Analysis Model

### Logistic Regression Model:

$$K/H = \beta_0 + \beta_1 * DUI + \beta_2 * Speed + \beta_3 * Fixed + \beta_4 * Head-On$$

Total records in our study Data set:

1,226

Records with missing values:

14 (1.1%)

July 14<sup>th</sup>, 2003

[www.kiprc.uky.edu](http://www.kiprc.uky.edu)

29<sup>th</sup> TRF 2003, Denver

## Results for the Gold Standard

Parameter	OR(95% CI)	Estimate
DUI	2.51 (1.58 3.98)	0.9189
Speed	1.58 (1.18 2.10)	0.4546
Fixed	1.70 (1.24 2.33)	0.5311
Head-on	1.70 (1.04 2.77)	0.5316

This Gold Standard result is used to compare with all other results.

**Conclusion:** comparing motorcyclists with DUI to motorcyclists without DUI, the odds of being killed or hospitalized are 2.5 times greater than the odds of not being killed or hospitalized, when other factors are controlled.

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## Imputation Model

### Analysis Model:

$$K/H = \beta_0 + \beta_1 * DUI + \beta_2 * Speed + \beta_3 * Fixed + \beta_4 * Head-On$$

### Imputation Model:

$$K/H \quad DUI \quad Speed \quad Fixed \quad Head-On$$

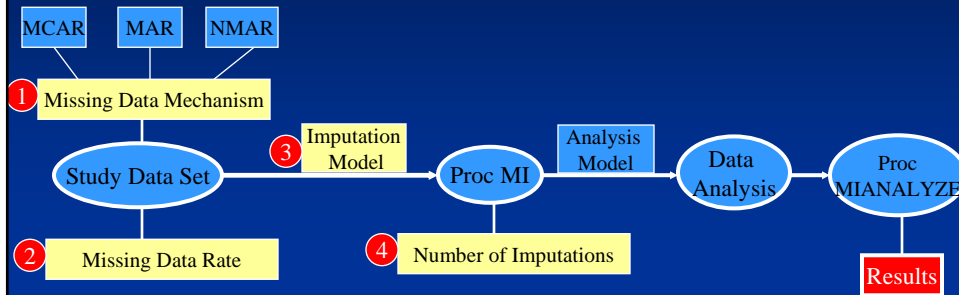
Note: The imputation model tells software how to impute the missing values, based on the relationships in the available data. The imputation model does not have to be identical to the analysis model, but at least it should include all of the analysis covariates. You can add any additional variables that are correlated to the variables that have missing values.

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

# SA: 1 Missing Data Mechanism



July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

# SA: 1 Missing Data Mechanism

- **Missing Completely At Random (MCAR)**
  - DFN: the missing data values are a simple random sample of all data values.
  - We simulated this condition by using SAS Proc SurveySelect to pick a random sample from the gold standard data set, then set *DUI = missing* for those selected cases.
- **Missing At Random (MAR)**
  - DFN: the probability of missing values on one variable is unrelated to the values of this variable, after controlling for other variables in the analysis
  - We simulated this condition by setting *DUI = missing* for riders aged 46 or older
- **Not Missing At Random (NMAR)**
  - DFN: the probability of missing values on one variable is related to the values of this variable even if we control other variables in the analysis
  - We simulated this condition by setting *DUI = missing* for uninjured riders who were not suspected of DUI (*DUI="NO"*).

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

Created 3 data sets from the gold standard data set with different missing data mechanisms, but with the same percent missing values for DUI (25%)

Parameter	MCAR 25% missing on DUI			MAR 25% missing on DUI			NMAR 25% missing on DUI		
	E	SE	P	E	SE	P	E	SE	P
Intercept	-1.7336	0.1096	0.0001	-1.7259	0.1092	0.0001	-1.7204	0.1092	0.0001
DUI	0.8544	0.2664	0.0016	0.8286	0.2623	0.0018	0.5791	0.2223	0.0092
Speed	0.5018	0.1449	0.0005	0.4843	0.1448	0.0008	0.4812	0.1443	0.0009
Fixed	0.4927	0.1610	0.0022	0.5079	0.1597	0.0015	0.5400	0.1578	0.0006
Head-on	0.5133	0.2485	0.0388	0.5133	0.2486	0.0389	0.5103	0.2475	0.0393

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

**Sensitivity analysis on missing data mechanism:**

Different

1 Missing Data Mechanism

Same

2 Missing Data Rate (25%)

Same

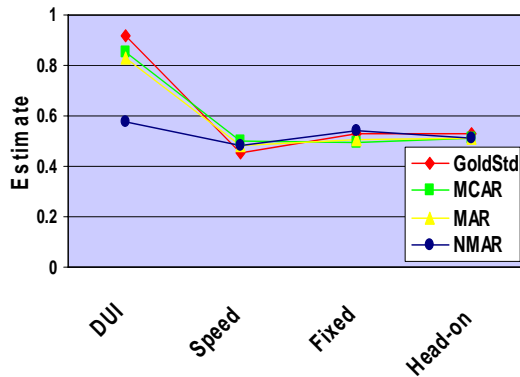
3 Imputation Model

Same

4 # of Imputations

What is the result?

Estimates for Parameters with Different Missing Data Mechanisms



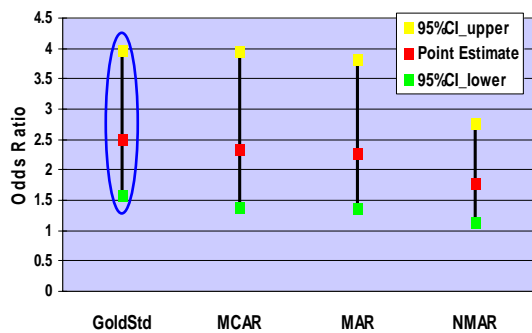
July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## Conclusions of SA on Missing Data Mechanism

Point Estimate and 95% CI for DUI with Different Missing Data Mechanisms



•Even if we used the simplest imputation model MI was able to produce results that are consistent with the Gold Standard when the missing data mechanisms were MCAR or MAR, but not NMAR

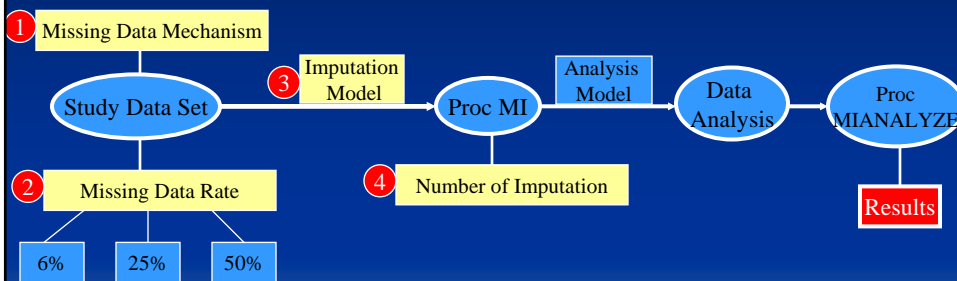
•we would predict the increased odds of death or hospitalization for riders suspected of DUI to be 1.78 (1.15 2.76) for NMAR, while our Gold Standard predicts it to be 2.51 (1.58 3.98).

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## SA: 2 Missing Data Rate



July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

SA: **2** **Missing Data Rate**

- **Data sets with MCAR** (Test on percentage of values missing for DUI as 6%, 25%, 50% respectively)
- **Data sets with MAR** (Test on percentage of values missing for DUI as 6%, 25%, 50% respectively)

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

Create 3 data sets with **MCAR** from the gold standard data set having values missing for DUI as 6%, 25%, and 50% respectively.

Parameter	MCAR 6% missing on DUI			MCAR 25% missing on DUI			MCAR 50% missing on DUI		
	E	SE	P	E	SE	P	E	SE	P
Intercept	-1.7361	0.1094	0.0001	-1.7336	0.1096	0.0001	-1.7377	0.1119	0.0001
DUI	0.9447	0.2429	0.0001	0.8544	0.2664	0.0016	0.8457	0.2973	0.0065
Speed	0.4812	0.1446	0.0009	0.5018	0.1449	0.0005	0.4831	0.1460	0.0009
Fixed	0.5213	0.1584	0.0010	0.4927	0.1610	0.0022	0.5200	0.1617	0.0013
Head-on	0.5245	0.2489	0.0351	0.5133	0.2485	0.0388	0.4936	0.2508	0.0490

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

Create 3 data sets with **MAR** from the gold standard data set having values missing for DUI as 6%, 25%, and 50% respectively.

Parameter	MAR 6% missing on DUI			MAR 25% missing on DUI			MAR 50% missing on DUI		
	E	SE	P	E	SE	P	E	SE	P
Intercept	-1.7382	0.1095	0.0001	-1.7259	0.1092	0.0001	-1.7502	0.1109	0.0001
DUI	0.9191	0.2334	0.0001	0.8286	0.2623	0.0018	1.2722	0.3298	0.0002
Speed	0.4836	0.1449	0.0008	0.4843	0.1448	0.0008	0.5063	0.1473	0.0006
Fixed	0.5076	0.1590	0.0014	0.5079	0.1597	0.0015	0.5234	0.1597	0.0010
Head-on	0.5174	0.2486	0.0374	0.5133	0.2486	0.0389	0.5371	0.2487	0.0308

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

### Sensitivity analysis on Missing Data Rate?

Same

① Missing Data Mechanism  
MCAR or MAR

Different

② Missing Data Rate

Same

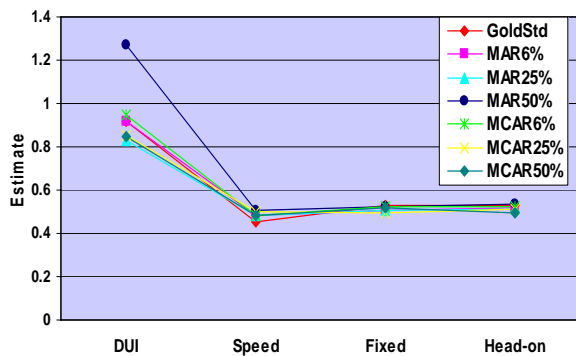
③ Imputation Model

Same

④ # of Imputations

What is the result?

Estimates for Parameters with Different Missing Rates

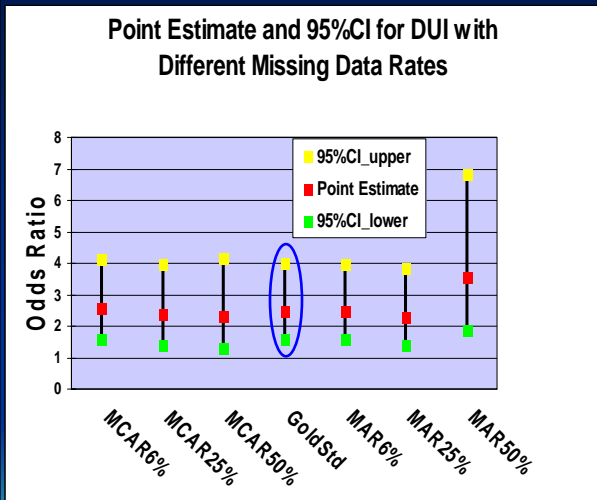


July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## Conclusions of SA on Missing Data Rate



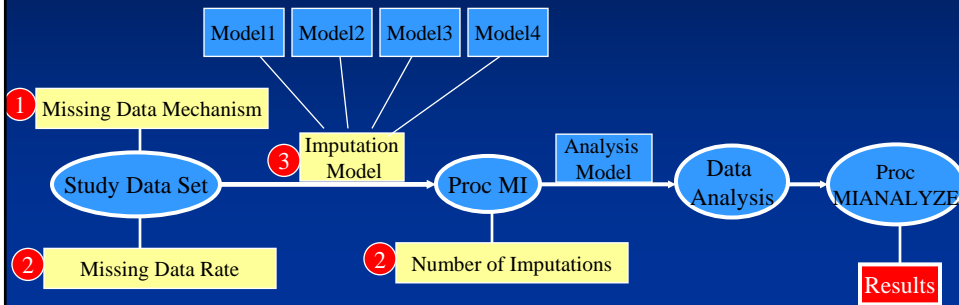
•It shows that the simplest imputation model is not sufficient to handle very high missing data rates with MAR. Later we will find out whether we can improve the MAR 50% results by changing our imputation model.

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## SA: 3 Imputation Model



July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## SA: **3** Imputation Model

- Data set with MAR and values missing for DUI=50%
- Tests on the following 4 Imputation models
  - Model1: D/H DUI Speed Fixed Head-on  
*Model1 = Analysis model, it is the simplest imputation model*
  - Model2: Model1 + age\_group + colltime (Categorical)
  - Model3: Model1 + age\_group + hour (Continuous)
  - Model4: Model1 + age\_group + hour\_normal (Continuous)  
*We are adding age and collision time to help predict DUI in Model2, Model3, and Model4*

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

Use 4 different imputation models to do MI on the same data set with MAR, 50% missing on DUI.

Parameter	Model 2 50% missing on DUI			Model 3 50% missing on DUI			Model 4 50% missing on DUI		
	E	SE	P	E	SE	P	E	SE	P
Intercept	-1.8110	0.1222	0.0001	-1.8081	0.1235	0.0001	-1.8034	0.1238	0.0001
DUI	1.0127	0.2948	0.0016	0.9814	0.2966	0.0024	0.9563	0.2813	0.0015
Speed	0.5079	0.1466	0.0005	0.5021	0.1463	0.0006	0.5081	0.1469	0.0005
Fixed	0.5370	0.1604	0.0008	0.5404	0.1601	0.0007	0.5371	0.1598	0.0008
Head-on	0.5554	0.2537	0.0286	0.5477	0.2552	0.0320	0.5561	0.2521	0.0274

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

### Sensitivity analysis on Imputation Model

Same

1 Missing Data Mechanism  
MAR

Same

2 Missing Data Rate (50%)

Different

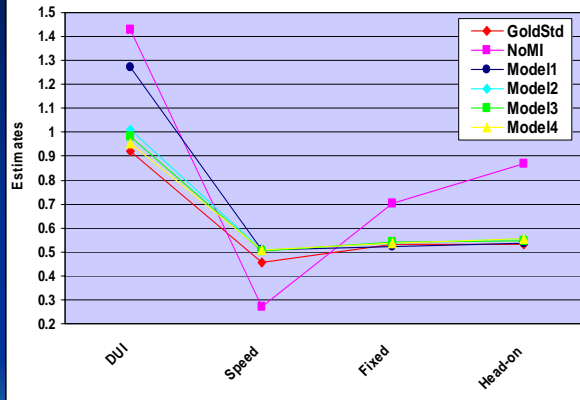
3 Imputation Models

Same

4 # of Imputations

What is the result?

### Estimates for Parameters with Different Imputation Models



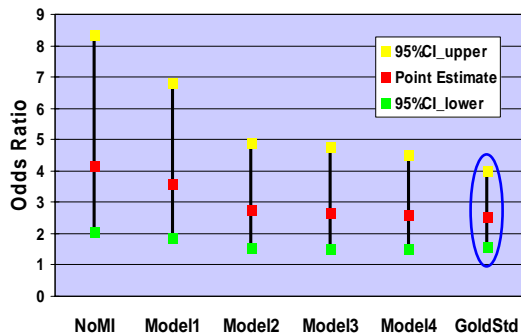
July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## Conclusions of SA on Imputation Models

### Point Estimate and 95% CI for DUI with Different Imputation Models



•Models 2, 3, and 4 are all improvements over model 1, and produced DUI parameter estimates and 95% CI widths close to those of the Gold Standard.

•So even with 50% missing values (MAR), we are able to get a good result by using a richer imputation model.

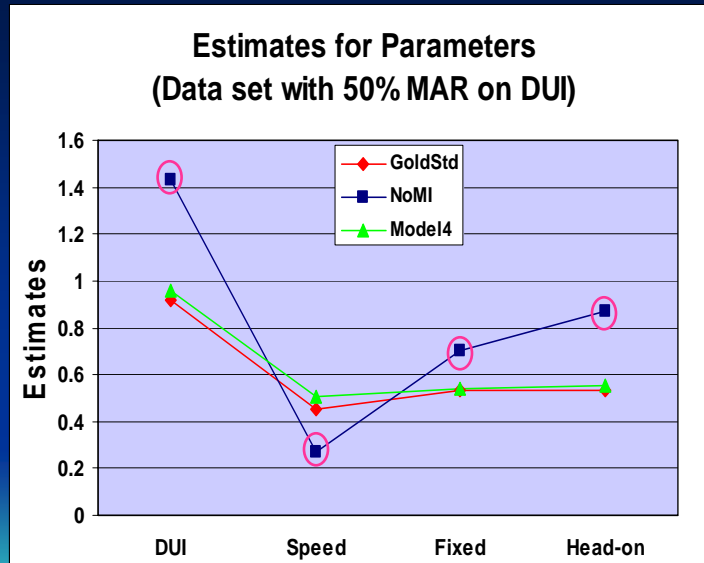
•The higher percent missing values (MAR) in your data set, the more you must include additional predictors in the imputation model.

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## Comparison of No MI and Model 4 to the Gold Standard

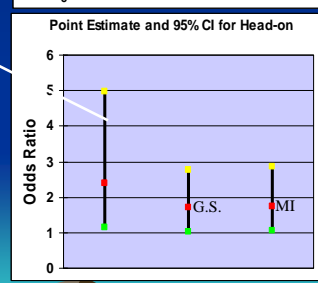
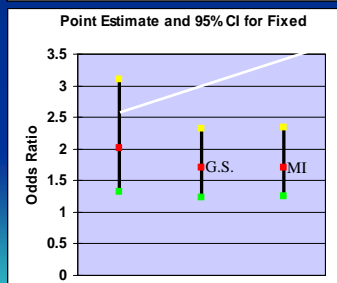
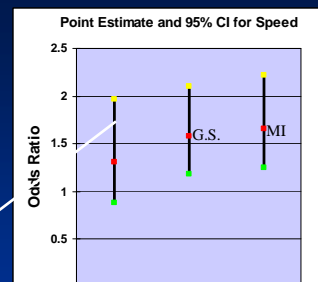
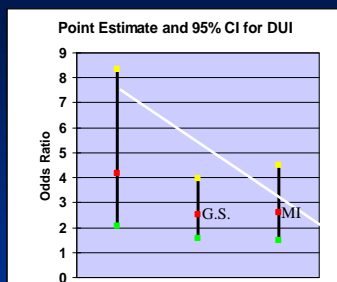


July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## Comparison of No MI and Model 4 to the Gold Standard



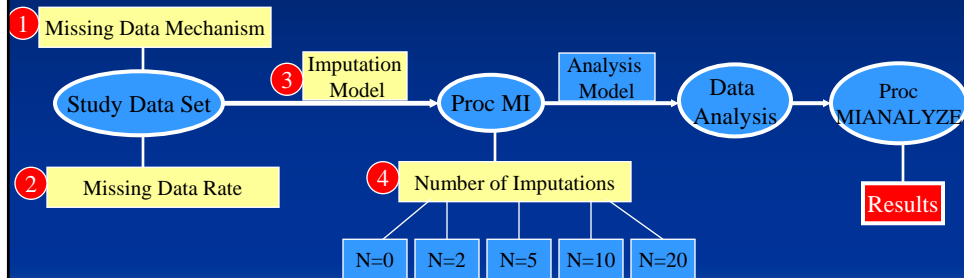
No MI

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## SA: 4 Proc MI: Number of Imputations



July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## SA: 4 Proc MI: Number of Imputations

- Data set with MAR and values missing for DUI=50%, use Model4 to do MI
- Test on different number of imputations
  - N=0
  - N=2
  - N=5
  - N=10
  - N=20

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

Use same imputation model (Model4), but different number of imputations to do MI on the same data set with **MAR**, 50% missing on DUI.

Parameter	N=5 50% missing on DUI			N=10 50% missing on DUI			N=20 50% missing on DUI		
	E	SE	P	E	SE	P	E	SE	P
Intercept	-1.7975	0.1177	0.0001	-1.8034	0.1238	0.0001	-1.7898	0.1204	0.0001
DUI	0.8658	0.2537	0.0023	0.9563	0.2813	0.0015	0.9942	0.3176	0.0026
Speed	0.4971	0.1457	0.0006	0.5081	0.1469	0.0005	0.5016	0.1465	0.0006
Fixed	0.5448	0.1610	0.0007	0.5371	0.1598	0.0008	0.5286	0.1599	0.0010
Head-on	0.5652	0.2522	0.0251	0.5561	0.2521	0.0274	0.5506	0.2509	0.0282

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

### Sensitivity analysis on Number of Imputations

Same

1 Missing Data Mechanism  
**MAR**

Same

2 Missing Data Rate (50%)

Same

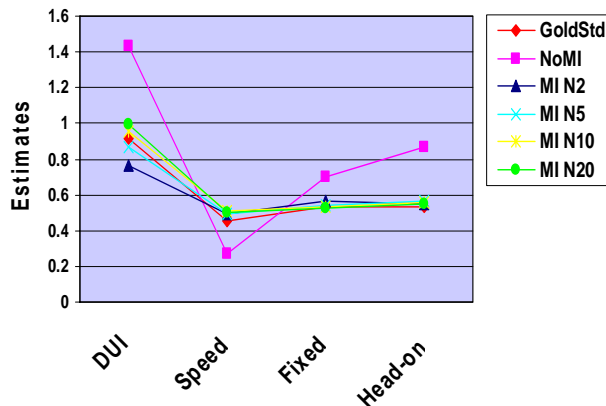
3 Imputation Model

Different

4 Number of Imputation

What is the result?

Estimates for Parameters with Different Number of Imputations



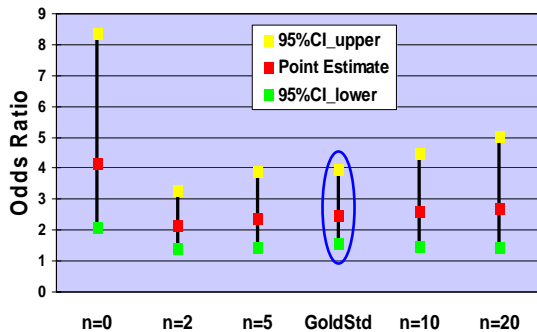
July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## Conclusions of SA on Number of Imputations

Point Estimate and 95% CI for DUI with Different Imputation Numbers



•In our example, n=5 to 10 is enough to get good results for data set with 50% MAR on DUI.

•No MI (complete cases only), we would conclude that: motorcyclists with DUI had 4.2 (2.1, 8.4) times more likely killed or hospitalized than motorcyclists without DUI. But from the Gold Standard, the OR is 2.5 (1.5, 4.0)

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver

## Summary---Answers?

- **May I use MI to deal with missing data problems for my data sets?**

Seems a good idea to try MI. Depend on the missing data mechanisms of variables with missing values in your data sets (however, even our results with MI for NMAR were better than No MI)

- **How can I believe that the MI will give me the better analysis results?**

We found that using MI on our example gave us much better analysis results than No MI (the complete cases only) **LIMITATION:** our example only included one variable with missing data. We haven't investigated the performance of MI in an application with missing values on several variables.

- **How can I get better analysis results by using MI?**

Understand the relationship between variables in your data sets;  
 Know the missing data mechanisms of variables;  
 Determine the percent of missing information;  
 Build a reasonable imputation model;  
 Use Proc MI options wisely

July 14<sup>th</sup>, 2003

www.kiprc.uky.edu

29<sup>th</sup> TRF 2003, Denver