Recent Developments in the Method of Estimating Missing Blood Alcohol Concentration (BAC) Values in FARS

Rajesh Subramanian and Dennis Utter

National Center for Statistics and Analysis

Presented by Rajesh Subramanian

National Center for Statistics and Analysis

Objective

Determine the extent of impaired driving in Fatal Motor Vehicle Traffic Crashes in the U.S.

Metric

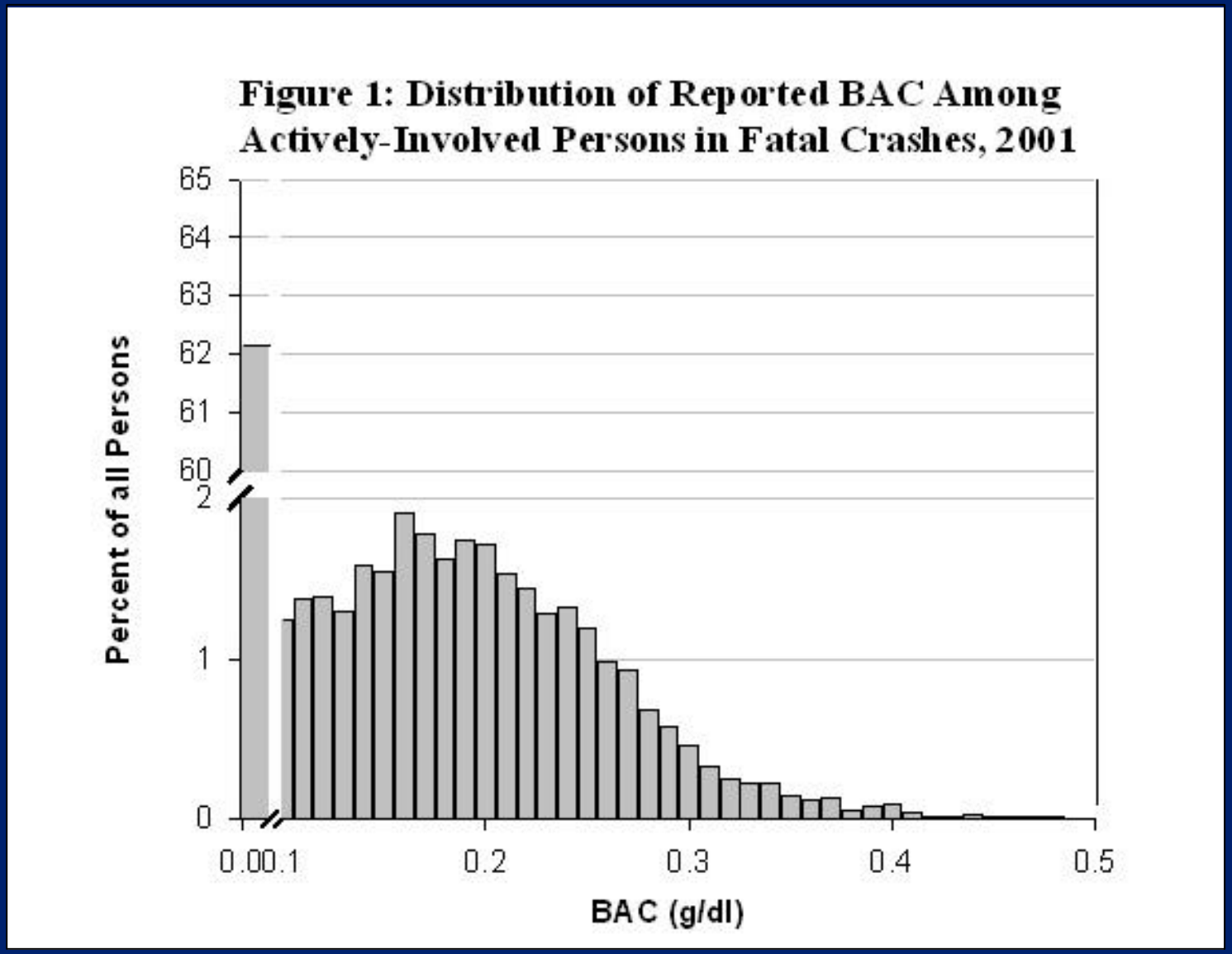
Alcohol-Related Fatality

A Fatality that occurs in a Motor Vehicle Traffic Crash where at least one of the involved Drivers, Pedestrians or Pedalcyclist has a Blood Alcohol Concentration (BAC) of 0.01 g/dl or above.

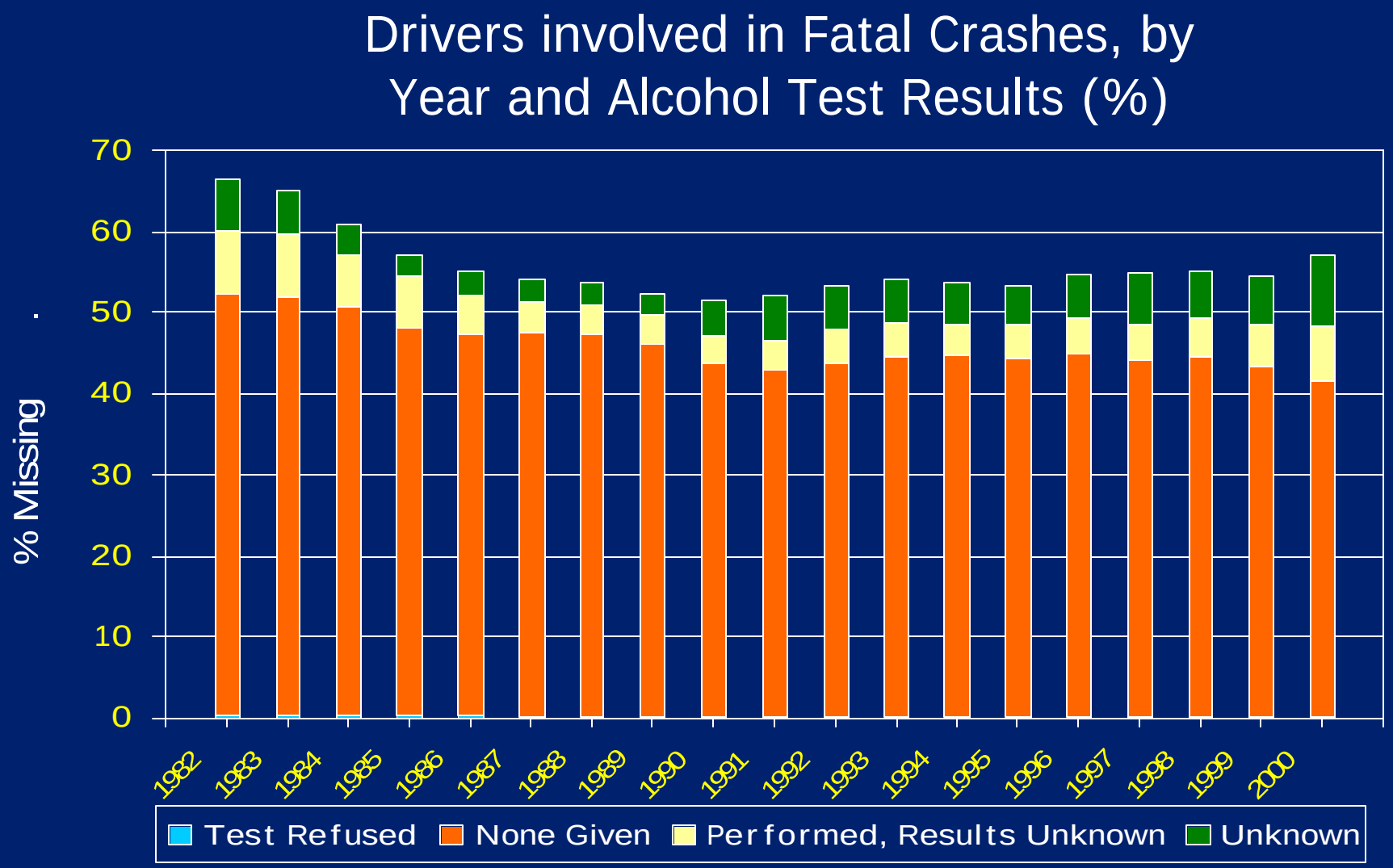
Data System

- Fatality Analysis Reporting System (FARS)
  - Census of all Fatal Crashes in the U.S.
    - About 38,000 Crashes in 2002.
    - Involves 58,000 Drivers.
    - Resulting in 43,000 fatalities.
- BAC reported to FARS
  - Breath test for Surviving Drivers.
  - Medial Examiner results for Fatal Drivers.

BAC – Semicontinuous Nature



The Problem: Missing BAC Values



Source: 1982-1999 Final FARS, 2000 Annual Report File

Effect of Missing BACs / Imputation

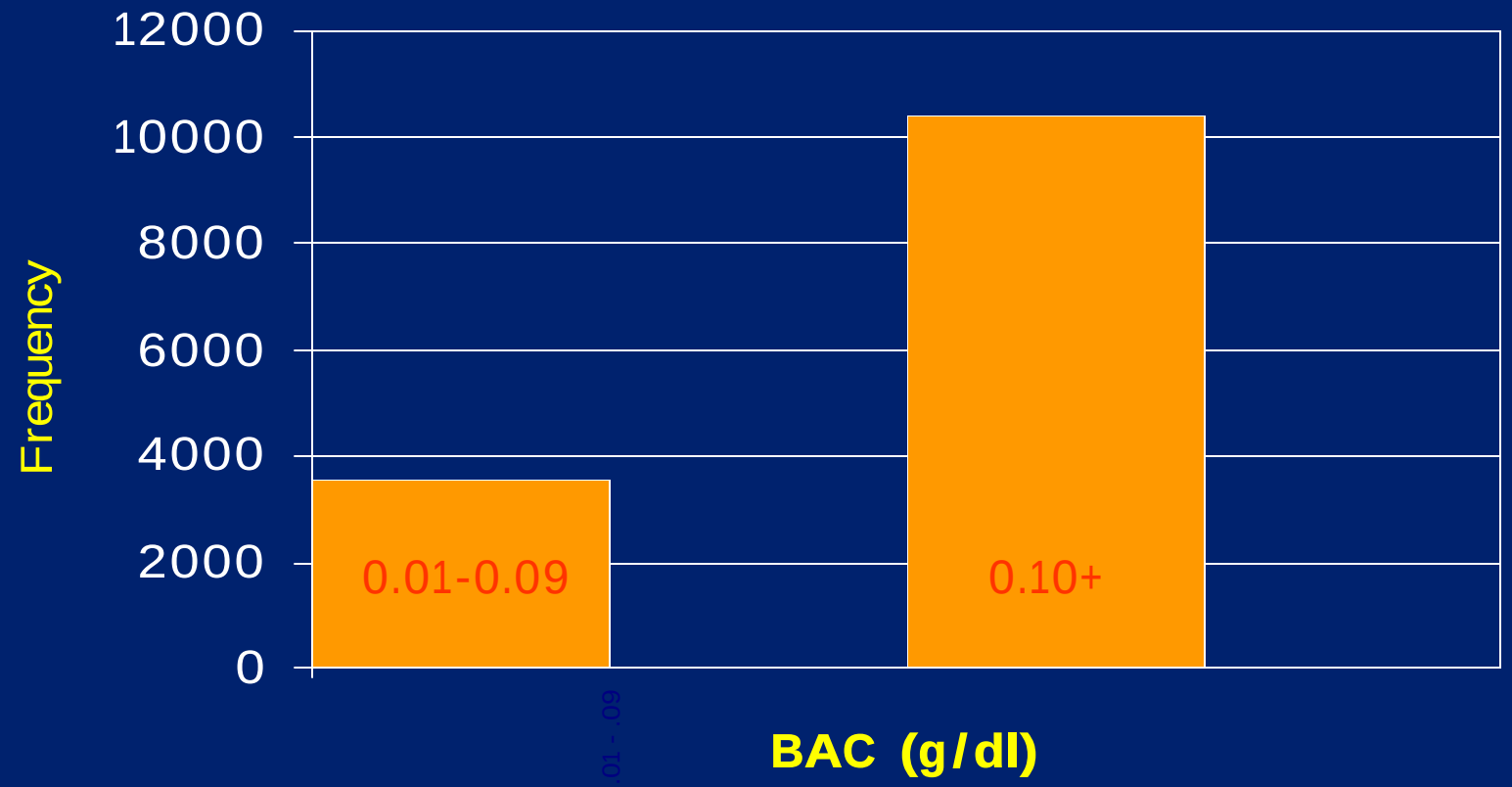
- Present National Statistics on Alcohol Related Crashes Based on All Cases
  - Imputing missing BACs provides a way to evaluate extent of alcohol involvement in all fatal crashes.
- Invalid inferences can be drawn on alcohol-related crash characteristics if based only on known BACs.

Prior Imputation Approaches

- Hot-decking
- Two-category Discriminant Analysis
  - Probability of Alcohol Involvement being in Yes/No categories.
- Most recently, three category Discriminant Analysis
  - Probability of alcohol involvement in one of three categories
    - (1) No Alcohol (2) BAC=0.01-0.09 (3) BAC=0.10+

Estimates from Previous Approach

Distributions of BAC for Drivers Involved in Fatal Crashes (excluding BAC=0), FARS



Source: 2000 Annual Report File

Disadvantages of Prior Approach

- Although BAC is continuous, reporting is along two three categories
  - No Alcohol, Low Alcohol (0.01-0.07), High Alcohol (0.08+)
- Impute probabilities instead of actual values.
- Cannot study problem along various BAC levels.

Multiple Imputation Approach

- Implementation of Multiple Imputation (Rubin, 1987) under the General Location Model or GLOM (Schafer,1997) in FARS.
- Implemented with 2001 data
- Revised estimates back to 1982 data – the first year NHTSA started reporting alcohol involvement.

FARS Variables used to Impute Missing BAC

Variables Used	
Police Reported Drinking	License Status
Age category	Previous Incidents (DWI, etc.)
Gender	Day of the Week
Use of restraint	Time of the Day
Injury severity	Vehicle Role
	Relation to Roadway

Overview of Imputation Process

- Step 1
  - Choose set of variables that are significant in predicting dichotomous BAC or BAC2 (BAC=0 vs. BAC≠0)
- Step 2
  - Conditional on case having non-zero dichotomous BAC, choose set of variables that are significant in predicting continuous BAC.
- Step 3
  - Combine the results from Steps 1 and 2 into one general model and impute missing BAC.

Two-Stage Model

- Two-stage mix of
  - A Normal Distribution and a Condensed Point Mass
- Semicontinuous observations $BAC_1,\ldots,BAC_n$ are recoded as two variables $(BAC2_i,z_i)$, i=1,…,n, where

$$BAC2_i = \begin{cases} 1 \; if \; BAC_{iy_i} \neq 0, \\ 0 \; if \; BAC_i = 0, \end{cases}$$

$$z_i = \begin{cases} g(BAC_i) \; if \; BAC_i \neq 0, \\ N/A \; if \; BAC_i = 0, \end{cases}$$

where g is some transformation (e.g., log) chosen to make positive BAC values most nearly normal, and the conditional distribution of zi when BAC2i=1 is assumed to be $N(m, s^2)$

Imputation Domains and Process Flow

- Steps 1 to 3 are performed within each vehicle class
  - Cars, Utility Vehicles, Minivans, Medium and Heavy Trucks, Motorcycles, Other.
  - Differing Driver Characteristics (e.g. Minivans vs. Motorcycles)
- Non-occupants are treated as a separate class
  - Lesser number of predictor variables (No license status, restraint use, etc.).

GLOM - General Location Model (Schafer, 1997) Step 1

- Dichotomous BAC (BAC2)
  - Linked to categorical covariates using loglinear model.
  - Simple association between BAC2 and each covariate.
  - Model selection by stepwise procedure
    - Begin with null model, test significance of each term not in model using likelihood-ratio test.
    - Most significant term entered into model – drop any covariate whose significance drops below the 0.1 level due to term addition.
    - Process repeated until all covariates in the model are significant at the 0.1 level and no covariates outside model is significant at 0.1 level.
- Obtain set of covariates significant in predicting BAC2.

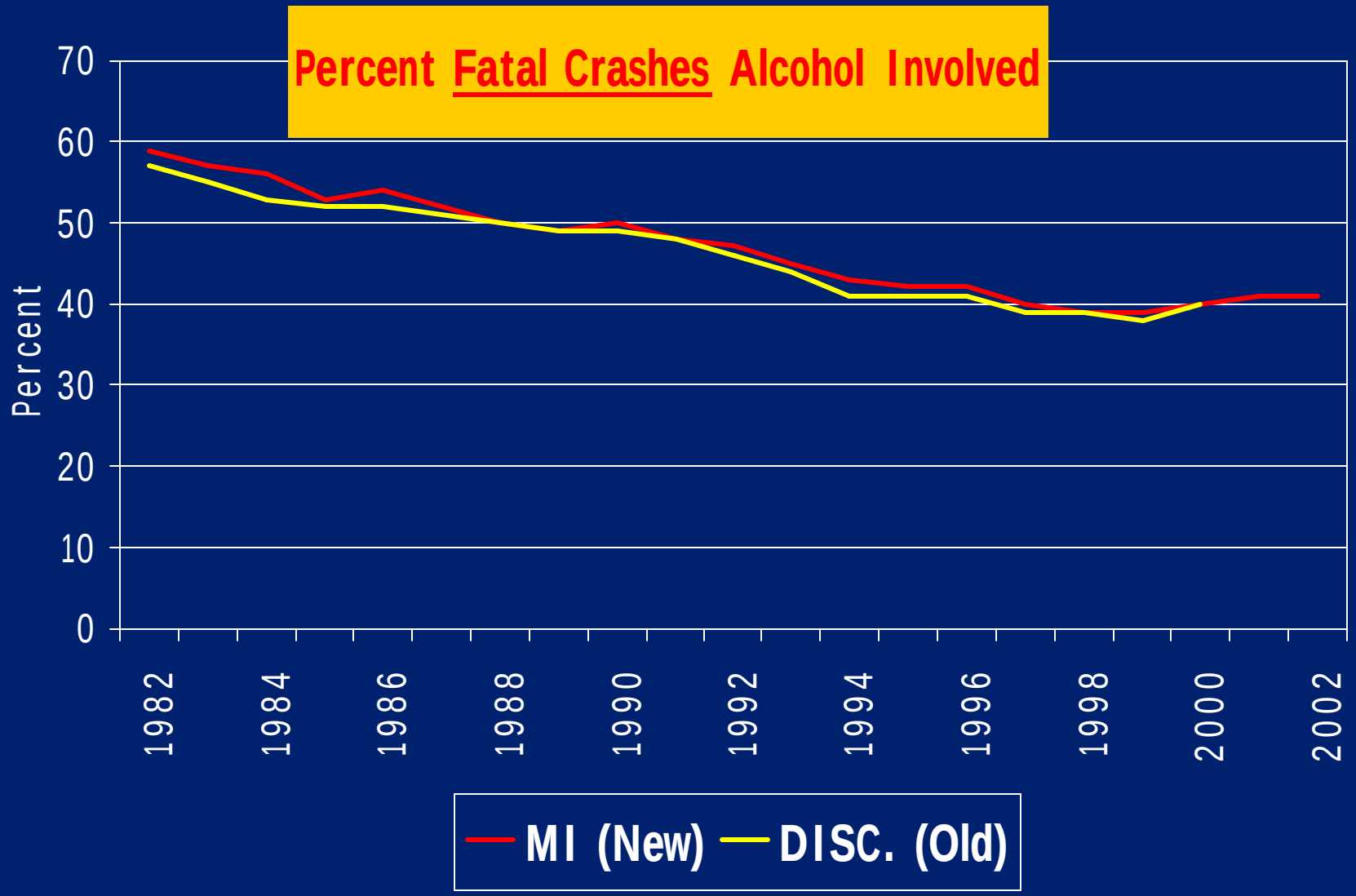
GLOM - General Location Model (Schafer, 1997) Step 2

- If BAC2=1
  - Linear regression model to associate positive BAC with covariates.
  - Choose $\log(BAC^{l})$ as $\log(BAC)$ imputes few implausibly high BACs.
  - Choose l based on ML method of Box and Cox.
    - Resulting ML estimate worked well for most vehicle classes.
    - Still produced implausibly high BACs for few vehicle classes.
- For each vehicle class, the transformation $g(BAC) = \log(BAC^{l+1})$ is chosen and significant predictors are chosen using least-square stepwise regression.

"Multiple" Imputations - General Location Model (Schafer, 1997)

- ML Estimates (MLE) of Model Parameters are found using ECM algorithm.
- Using MLE as seed, new Parameters are simulated from their posterior distribution using MCMC.
- Repeating process 10 times results in ten imputations of g(BAC)
- Inverse Transformation g^{-1} converts back to actual BACs.

Comparison of Estimates of Alcohol-Related Crashes in the U.S., 1982-2002



Validation Exercises:
2002 Preliminary to Final Data Changes

Comparison of Imputed (Preliminary) and Reported BAC values (Final) among Drivers involved in Fatal Crashes			
BACs	Preliminary (Imputed)	Final (Reported)	Percent Conforming
0	1,162	1,201	97%
>0	566	527	93%
Total	1,728	1,728	-

Variation by Police-Reported Drinking