Methodology Used To Create "Initial Visit" for
2005 National Hospital Ambulatory Medical Care Survey
Emergency Department Visit Records

I. Background

Both the National Ambulatory Medical Care Survey (NAMCS) and the National Hospital Ambulatory Medical Care Survey (NHAMCS) are currently conducted annually. The majority of publicly available data items are consistent year to year, however in any given year, certain items may be added, removed, or modified, based on various considerations such as demand, reliability, cost (for new items), etc.

From 2001 to 2004, a data item called "Episode of Care" (EPISODE) which included the values "Initial visit for problem", "Follow-up visit for problem" and "Unknown", was included to the Emergency Department (ED) component of NHAMCS. This item was not included in the 2005 and 2006 survey; however it will be re-introduced in the 2007 survey.

For researchers interested in the episode of care for visits in 2005, a new, imputed variable (INITVIS) was created using data from previous years, and is available on the 2005 public use file. It should be noted however, that because this is an imputed variable, it is not comparable to the EPISODE variable from previous years and should therefore be used with caution and only with appropriate footnotes in any year-to-year trend analysis (see sections III and IV below).

II. Methodology

The basic method for imputing INITVIS was to use previous data which included the EPISODE variable to fit a logistic model to predict the odds of an initial visit for each visit. The results of the regression (the betas for each variable in the model) were then "applied" to the 2005 data so that each record would have a value for INITVIS based on the probability of that visit being an initial visit. The data years used to create the model were 2003 and 2004 since these years included the mode of arrival variable which was highly correlated with initial visits.

Because of the differences in characteristics between visits to the ED for injuries versus visits for illnesses, two separate models were used to fit the data (one for injuries and one for illnesses). The methodology that follows was the same for each model.

A. Variables in the models

Only variables which showed a significant association ($p<.05$) with the odds of an initial visit were included in the models.

1. The injury model included the following variables:

Arrival by ambulance
Disposition of "Return/refer to other physician for follow up"
General type of injury/poison (fractures, sprains, etc.)
Immediacy with which patient should be seen
Left before being seen
Mechanism of unintentional injury (falls, machinery, etc.)
Patient age
Patient race/ethnicity
Patient seen in the last 72 hours
Presenting level of pain

2. The illness model included the following variables:

Disposition of "No follow up"
Disposition of "Return/refer to other physician for follow up"
Immediacy with which patient should be seen
Major disease category (infectious, digestive, etc.)
Patient age
Patient race/ethnicity
Patient seen in the last 72 hours
Presenting level of pain
Principal reason for visit module (symptoms, disease, diagnostic, etc.)
Specific symptom (general, referable to respiratory, etc.)

B. Creating and assigning probabilities to the 2005 data

For each visit record in the 2005 data, a probability that the visit was an initial visit was calculated by summing the intercept and the betas from the appropriate regression model for each of the covariates in the record (note – it was actually odds that were calculated since a logistic regression model was used, but the odds were converted to probabilities). For example, if the only information in a visit record was the patient age and race/ethnicity, then the probability that the visit was an initial visit would be the sum of the intercept and the betas for the patient's age group and race/ethnicity group.

In order to assign a value for INITVIS and introduce variability in the data (so that visits with similar characteristic are not all treated exactly the same), a random number from 0 to 1 was generated for each visit record. If the probability of an initial visit was greater than the random number generated, the imputed value for INITVIS was set to 1; otherwise it was set to 0.

C. Testing the methodology

As stated above, the models were fit to two years of data (2003 and 2004), and then "applied" to data from 2005. Since the EPISODE variable was not included in 2005, there is no way to test the accuracy of the imputed values against the true initial visit values. For this reason, we tested the entire methodology described above by fitting a model to data from 2003 and "applying" the results from the regression to 2004. We then compared the imputed values from the models (INITVIS) with the values in the EPISODE variable. We also did the reverse (fit the model to 2004 and applied it to 2003), and in each case there was a high correlation between INITVIS and EPISODE. The resulting correlations are shown in the table below. We expect that by fitting the model to two years of data rather than only one, we will get an even better estimate of the true episode of care.

Correlation Table

Model	Year[1]	Applied[2]	Correlation[3]
Injury	2003	2004	0.87
	2004	2003	0.76
Illness	2003	2004	0.93
	2004	2003	0.86

[1] The year of data used to create the model.

[2] The year of data to which the results of the model were applied.

[3] The correlation between the actual % of initial visits and the estimated % based on the model. The variables used in the correlation calculations were age, sex, race/ethnicity, and type of injury for the injury model; and age, sex, race/ethnicity, and major disease category for the illness model.

III. Limitations

While the EPISODE variable has 3 values (initial visit, follow up, and unknown), the INITVIS variable only has 2 (initial visit and not initial visit). A binary outcome of initial visit was the only response predicted since roughly 90% of ED visits from 2001 to 2004 were initial visits, 5% for follow up, and 5% were unknown.

In addition, only the point estimates (not the variance) from the regression models were used in the probability calculations. However, random error was introduced as explained in the methodology above.

Further, as opposed to using information from the current year to impute missing values for the same year, data from previous years (2003-04) was used to impute values for a later year (2005). Therefore any differences in the factors influencing the likelihood of an initial visit which vary from year to year would have been unaccounted for. This is most likely not a significant issue, however, given the homogeneous pattern of the EPISODE variable from 2001 to 2004.

IV. Suggested Footnote

If the data are used to trend across years, an appropriate footnote should be included such as:

"Initial visit values for 2005 were based on a regression model-based imputation strategy using data from 2003-04. The item was not directly collected in 2005. For more information, go to www.cdc.gov/nchs/namcs.htm."