Mar 2007
Vol 4, Issue 3

Center for Mental Health Services Research
University of Massachusetts Medical School

CMHSR

Issue Brief

Estimating Frequencies from Multiple Source Data

Elizabeth Aaker, B.A., Charles Lidz, Ph.D. & Lorna Simon, M.A.

Consider the following: a psychiatric epidemiologist studying the prevalence of psychopathology in children based on reports from a number of sources for each child; a mental health services researcher estimating service usage from multiple provider databases; or a sociologist estimating the number of violent events based on interviews with both subjects and collaterals. These exemplify classic research situations in which investigators attempt to obtain epidemiological estimates from more than one source.

In contemporary statistics, the use of multiple source data to estimate frequency or prevalence is an evolving methodological process. In fact, several groups of researchers[1,2,3] are presently working to refine current models and improve the accuracy of information gleaned from multiple source data. Recently, CMHSR investigators developed a statistical model that estimates both the frequency of events and the prevalence in a population based on multiple source data.[4] This brief describes this approach for estimating the frequency of past events to enable researchers to integrate the model into their work and it suggests strategies for further testing of the model.

An Example

The following example demonstrates how the model can estimate the number of times a child (Sue) stopped to buy candy on the way to school. The data come from John (the candy store clerk) and Sue who both report the specific dates when Sue bought candy during the school year. In this case, the model combines the reports of Sue and John to estimate the frequency of Sue's candy buying. There are only two sources of data, the Subject (Sue) and the Collateral (John), and they only give positive reports (dates that Sue bought candy). Negative reports are all the dates that the sources did not report an incident because they did not remember Sue purchasing candy on those dates (See Table 1 on next page).

The Models

When developing a model to estimate frequency, it is important to consider two types of error. First, there could be times that both sources fail to report an event happening, when in fact, it did happen; these are *false negative reports.* False negative reports fall into Cell D in Table 1. The opposite is also true. There could be times that sources report an event happening, when in fact it did not happen; these are *false positive reports.* False positive reports could show up in Cells A, B or C. The model developed by CMHSR attempts to decrease the number of false negatives. The team also studied some of the influence of false positives on the model's estimates; however, the theory behind the model assumes no false positive reports.

A common problem in traditional methods of multiple source data frequency estimates is under-reporting, that is, failure to account for false negatives. For example, the traditional model adds up all positive reports by either the subject or the collateral (A+B+C). That is to say, it takes all the times that either Sue or John reported Sue stopping and it adds them together. It assumes that all incidents are reported at least once by either the subject or the collateral and that none are missed by any source.

The CMHSR alternative method uses the level of disagreement between the sources of data to adjust the estimates of frequency. Since it is unlikely that both the subject and the collateral will report all the events, even when their reports are combined, there are most likely some false negatives missed by the traditional model. The CMHSR Model, based on prior work[3], includes a correction factor

Table 1

	Subject (Sue) reports event happening (+)	Subject (Sue) does not report event happening (-)
Collateral (John) reports event happening (+)	**A** **15** dates both John and Sue remember	**B** **35** dates that John remembers but Sue does not
Collateral (John) does not report event happening (-)	**C** **35** dates that Sue remembers but John does not	**D** All the dates not mentioned by both Sue and John

to account for some of the unreported events. The equation for this model is A+B+C + (BC)/A. The correction factor, (BC)/A, is the product of the number of events in which the subject and collateral disagree divided by the number of events in which they agree. The CMHSR Model assumes that the sources are independent of each other. That is, no source is dependent on another source and all events have an equal likelihood of being reported by either source. In this example, it is reasonable to assume that the dates John and Sue reported are not dependent upon each others reports since we did not ask them within each others hearing. In addition, they were equally likely to report any of the events. Thus, the sources in this example are independent.

Previous research shows that the CMHSR Model catches more false negatives and seems to be less susceptible to false positives. More specifically, in this example, the traditional model estimates that Sue stopped 85 times (15+35+35) while, the CMHSR Model estimates 167 times (17+30+28 + (35*35)/15). This means that the CMHSR Model catches at least 82 incidents missed by the traditional model, thereby decreasing the number of false negatives. Researchers developed the CMHSR Model to primarily address error from false negative reports. However, when empirically tested CMHSR investigators found their model's estimates to be relatively robust with regard to false positives. The estimates of the CMHSR Model varied only slightly, while the traditional model's estimates varied considerably when the data changed to reflect different numbers of false positives.

Future Directions for Model Development

As shown through the CMHSR Model, the general technique of using the level of agreement to adjust the estimated frequency of events has much broader utility, including the possibility of determining risk factors for problematic conditions and behaviors. However, testing and development of the model is not complete. Researchers should continue to explore several directions to ensure the accuracy and superiority of the model. These directions include:

- Examining how the model might account for a potential relationship between the rate of false negatives and independent variables by splitting samples and applying the basic model to different subgroups;
- Testing the impact of non-random false positive reports to determine if, unlike random false positives which have no impact, there are circumstances in which non-random false positives may influence the model's estimates;
- Expanding the testing of the model to analyze data from more than two sources;
- Investigating the problem of missing data within the model by assessing the effect of various approaches to imputing missing values.

References

1. Horton, N.J., & Fitzmaurice, G.M. (2004). Tutorial in biostatistics: Regression analysis of multiple source and multiple informant data from complex survey samples. *Statistics in Medicine, 23,* 2911-2933.

2. Daskalakis, C., Laird, N.M., & Murphy, J.M. (2002). Regression analysis of multiple-source longitudinal outcomes: A "Stirling County" depression study. *American Journal of Epidemiology, 155,* 88-94.

3. Lie, R., Heuch, I., & Irgens, L.M. (1994). Maximum likelihood estimation of the proportion of congenital malformations using double registration systems. *Biometrics, 50,* 433-444.

4. Lidz, C.W., Banks, S., Simon, L., Schubert, C., & Mulvey, P. (2007). Violence and mental illness: A new analytic approach. *Law and Human Behavior, 31,* 23-31.