

## **Predicting Vulnerability of Indian Women to Domestic Violence Incidents**

**Debarchana Ghosh\***  
University of Minnesota

### **Abstract**

International events namely United Nations conferences on population and development, the Declaration on the Elimination of Violence Against Women, Platform of Action for United Nations and Beijing World Conference on Women recognize violence against women as a violation of basic human rights, impediment to women's autonomy and adverse repercussion on reproductive health. They further acknowledge that lack of data, statistics, and techniques to quantify incidents of violence against women make monitoring and surveillance of quick-response and preventive programs challenging. This paper attempts to address this issue by building models using modern data mining techniques namely classification tree and random forest to predict vulnerability of ever married women of age 15 to 40 years to domestic violence incidents in India. The data used in this study is obtained from the National Family Health Survey conducted in India during 1998-99.

---

(\* **Debarchana Ghosh**, PhD. Student. Department of Geography, University of Minnesota. E-mail: [ghos0033@umn.edu](mailto:ghos0033@umn.edu))

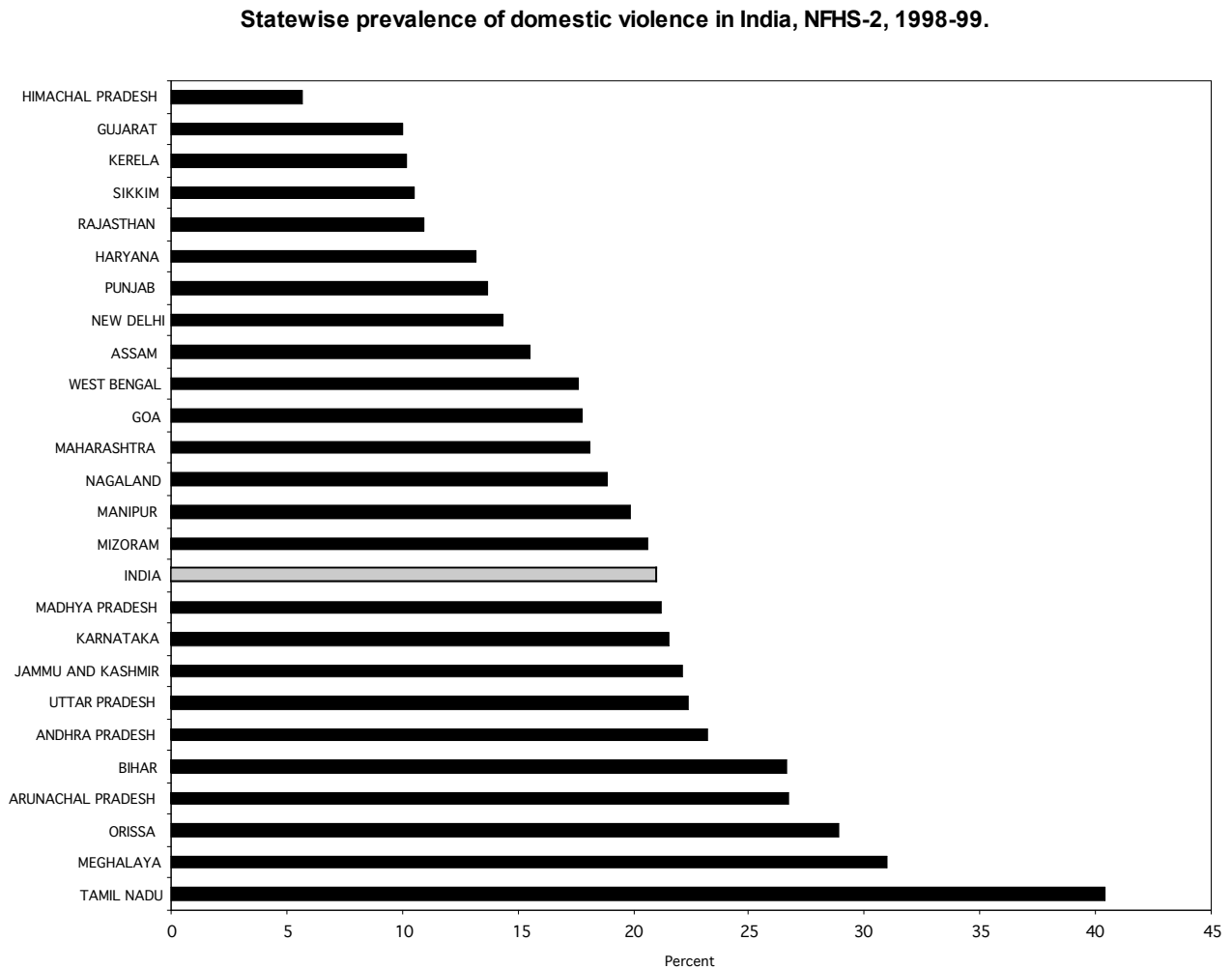
### **Introduction**

Women in India experience violence in various forms throughout their lives, and it cuts across boundaries of caste, class, religion, and region (Bhatti 1990; Daga 1998; Miller 1999; Mitra 1999; Rao 2000; Visaria 1999; Vindhya 2000). Among the different types of gender-based violence, domestic violence is the most common type prevalent in India. According to the National Family and Health Survey-2 (NFHS-2), 21 percent of ever-married women in India have been physically mistreated by their husbands, in-laws or other members of the household since the age of 15 years (IIPS 2000). Statistics from the National Crime Records Bureau show significant increase in reporting of gender-based violence from 31 percent in 1995 to 45 percent in 1999 (NCRB 1999). Evidences from population-based surveys suggest that between 21 to 48 percent of women from different socio-cultural settings in India have experienced domestic violence (INCLLEN 2000; Jejeebhoy 1998; Verma 2003; Visaria 1999). In another study of 4000 women reporting physical violence, 63 percent reported the experience more than three times (INCLLEN 2000). These statistics indicate that domestic violence in India is rarely an isolated event.

Further analysis of the prevalence rates of domestic violence incidents reveals statewide variation in India (Figure 1.) Tamil Nadu shows the highest prevalence with 41 percent of the women reporting domestic violence incidents since the age of 15 years. Andhra Pradesh, Karnataka, Meghalaya, Arunachal Pradesh, Mizoram, Orissa, Bihar and Jammu and Kashmir have prevalence rates higher than 20 percent. Himachal Pradesh shows the lowest prevalence of 5.8 percent, followed by Kerala (10.1 percent) and Gujarat (10.2 percent) (IIPS 2000).

**Figure 1. Prevalence of Domestic Violence in India, NFHS-2, 1998-99.**

Source: IIPS 2000



Prior studies on domestic violence have indicated that it is pervasive and deeply rooted in socio-cultural norms (Bhatti 1990; Daga 1998; Miller 1999; Mitra 1999; Rao 2000; Visaria 1999; Vindhya 2000). Kishor (2004) indicated several socioeconomic and cultural risk factors of domestic violence in her ‘multi-country’ empirical study of prevalence of domestic violence. However, little research has been conducted on the prediction of *vulnerability* of women to the experience of domestic violence incidents in the context of her socioeconomic and cultural background. This will help in effectively identifying the vulnerable target groups and sub groups of women who are at risk to experience domestic violence This is also important in protecting

women from domestic violence at the individual, family and community level by social workers, health officials, NGO's, police, and legal institutions. This paper attempts to address some of these issues.

### **Objective**

There are two main objectives of this paper. First, predicting vulnerability of an ever-married woman of age 15-49 years in India to domestic violence using classification tree and random forest algorithm. Second, to identify the most important risk factors associated to the experience of domestic violence.

### **Organization of the paper**

The paper is divided into four sections. The conceptual framework section discusses the context of occurrences of domestic violence with women's own background characteristics, their spouse, marital union, and characteristics of their household as potential risk factors. The next section explains the research design, the data source, variables included in the study, and the methodology. The 'analysis' section explains the predictive models of logistic regression, classification tree and random forest. The conclusion section, along with summarizing the findings also provides policy implications to combat the incidents of domestic violence incidents.

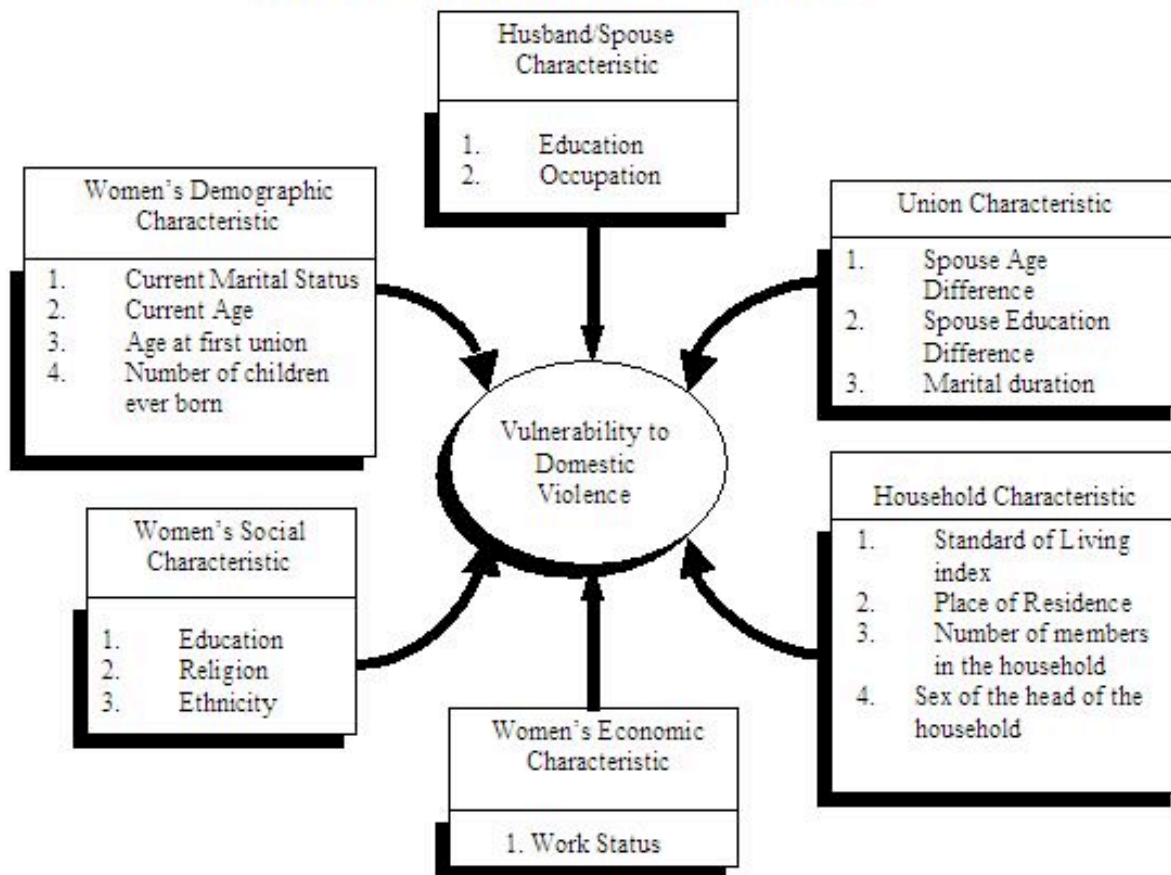
### **Conceptual Framework**

The causal factors and processes associated with the phenomenon of domestic violence are not clearly understood (Kishor 2004). However, by examining the selected background characteristics of the individuals and their relationships affecting spousal differences it is possible to tease out certain factors that are associated with an increased vulnerability of experiencing domestic violence.

Figure 2 shows the conceptual framework to predict the vulnerability of a woman to be physically mistreated by a perpetrator. The risk factors are categorized into six subdivisions. First category is women's demographic characteristics including current marital status, current age, age at first marriage and number of children ever born. Second, women's social

characteristics consists her education, religion, and ethnicity. Women’s economic characteristic is the third category. Fourth, women’s household characteristics include standard of living index, place of residence, number of members of the household, and sex of the head of the household. Fifth category, union characteristics includes spousal age and educational difference and marital duration. Finally, husband or spousal characteristics consist their education and occupation. The following paragraphs justify the inclusion of the above-mentioned potential risk factors for women to experience domestic violence in India based on the literature review.

Figure 2. A Conceptual Framework for the predictive models



### **Women's demographic characteristic**

**Current marital status:** The results of the multi-country study based on comparable Demographic and Health Survey (DHS) data indicated that prevalence of ever-experience of domestic violence varies among groups of women, namely, currently married and have been married only once, women who are married and have been married more than once, divorced or separated, and widowed women. Given that spousal violence is a common reason for divorce, it is not surprising that in most countries namely Cambodia, Columbia, Dominican Republic, Egypt, Haiti, India, Nicaragua, Peru, and Zambia, the study indicated that the highest rates of ever-reported domestic violence by women who are currently divorced/separated or in second or higher order marriage.

**Current Age:** Various studies have shown that women's age affects the likelihood that she would experience domestic violence (Daga 1998; Visaria 1999). Ever-experience of violence is generally hypothesized to increase with age, since older an ever-married woman is the longer has been her period of exposure to the risk of violence. However, the relationship of women's age and experience of domestic violence is not simple. It does not increase monotonically and fluctuates inconsistently within a narrow range of age (Kishor 2004).

**Age at first union:** A women's young age at first union is generally thought to be another risk factor for the experience of domestic violence (Kishor 2004). This hypothesis has both contextual and individual-level explanations. At the contextual level, age at marriage is a reflection of status of women (Mason 1987). Violence is often positively correlated with very early marriages in societies where women's status is low. At the individual level, a woman's age at marriage is related to her risk of experiencing violence because when she marries at younger age she has probably not given a chance to acquire understanding and maturity needed to ensure her security in marriage.

**Number of children ever born:** Several studies have also shown that the vulnerability of experiencing domestic violence is positively related to the number of children (Ellsberg 2000; Martin 1999). The association between violence and number of children could be conceptualized such that when there are more children in a household there is economic insecurity, insufficient

resources, which may lead to disturbing levels of stress for the head of the household. This in turn may further lead to violence in some instances. Hence more the number of children, the greater is the likelihood of violence (Martin 1999). On the other hand, the presence of greater numbers of children in a household might be a result of spousal violence rather than a cause (Johnson 2003). That is, women who are subject to partner violence may be less able to control their own sexuality and fertility than women who are not subject to violence do. Thus, the direction of the relationship between number of children and domestic violence remains unclear.

### **Women's demographic characteristic**

**Education:** Education has been one of the sources of empowerment for women. It has given women the ability to gather and assimilate information, manipulate and control the modern world, secure and protect themselves from any form of violence (Malhotra 1997; Kishor 2000, 2004). It is hypothesized that women with more education have greater abilities to protect themselves in times of need, such as when dealing with a violent partner. Thus, it is expected that women with higher levels of education experience less violence. However, it is also speculated that there may be a transition for women who have begun their autonomy. For example, the urban, better educated, and economically active women may in fact suffer more violence than other women precisely because of the greater agency they exert in their own lives, thereby challenging existing gender norms (Daga 1998; INCLEN 2000).

**Religion:** Several studies have shown that women's religion affects the likelihood that she would experience domestic violence (Daga 1998; Visaria 1999). However, we cannot hypothesize specific effects of religion on the likelihood of domestic violence.

**Ethnicity:** Various studies have also shown that women's ethnicity has some association with the risk of experiencing domestic violence (Daga 1998; Visaria 1999). In the socio-economic hierarchy of India, scheduled caste, scheduled tribes and other backward classes are the most deprived and disadvantaged groups. Usually they have higher number of children, low per capita income, and insufficient resources, which may lead to exacerbated levels of stress for the head of the household and which in turn may lead to violence in some instances.

### **Women's economic characteristic**

**Work Status:** Economic independence is also one of the main sources of women's empowerment. The relationship between work status of a woman and her risk to experience domestic violence can be conceptualized in a way that women who are engaged in paid employment have more say over financial and household matters than women who are not active in the labor market (Malhotra 1997; Garcia 2000). Thus, women who are currently employed are at lower risk to experience domestic violence. However due to the transition phase towards autonomy mentioned earlier, the changing economic control from men to women can also lead to more incidents of violence for women.

### **Husband or partner's characteristic**

To fully understand domestic violence, the background characteristics of husband or partner who is the alleged perpetrator of the violence also needs to be examined.

**Husband's Education:** According to a multi-country study of prevalence and incidence of domestic violence it can be hypothesized that the relationship between husband's education and violence is negative and monotonic (Kishor 2004). However, this association is also subject to fluctuation and inconsistencies. For example, in case of Haiti, the relationship between education and violence is positive and monotonic (Kishor 2004).

**Husband's Occupation:** Some of the literature indicates that in developing societies where exclusively sons inherit agricultural land, women are more likely to be culturally devalued (Dyson 1983) and hence at a higher risk of violence. However the relationship of husband's/partners occupation and domestic violence is not very clear in the contemporary literature.

### **Union characteristics**

Risk factors for women's vulnerability to spousal violence include not only their own and their husband's background characteristics individually, but also to the extent how these socioeconomic and cultural characteristics are compatible with each other (Kishor 2004). Status inconsistency theory when applied to the marital issues suggests that when two people of

incompatible ascribed or achieved status engage in a marital union it may result in tensions and further lead to marital dissatisfaction (Mueller 1979).

**Spousal age difference:** Wide differences in spousal age, in which the husband is much older than the wife, are hypothesized to imply power imbalances in the relationship. Combination of seniority (achieved) and masculinity (ascribed) in many cultures puts wives younger than their husband at a comparative disadvantage position (Kishor 2004). However, there is little evidence in the empowerment literature regarding the effect of converse situation where the wife is older. But Kishor (2004), in her multi-country profiling of domestic violence was of the opinion that it may be more likely that because relationships in which women are older than their husband are so contrary to the normative marital arrangement in most societies, they may be at greater risk for marital disharmony.

**Spousal educational difference:** The literature suggests that men with higher educational status than women having both higher ascribed (on the basis of gender) and achieved (on the basis of higher educational attainment) status are more likely to assert unequal, and even violent power in the relationship (Hornung 1981). It has also been suggested from various other studies that when women have greater achieved status than their husbands, there is an increased vulnerability of marital discord (Hornung 1981; Daga 1998; INCLEN 2000).

**Marital duration:** The rate of ever-experience of domestic violence is expected to rise with marital duration because a longer marriage provides a greater period of exposure to the event of violence. However, this relationship could also be argued in converse manner. Marital duration is considered a proxy for compatibility in a marriage, particularly in cultures where divorce is legal and socially accepted. In such cases, the experience of violence is likely to be negatively associated with marital duration (Kishor 2004).

### **Household characteristic**

An important context of women's life is the characteristics of the household in which they reside including the location of the household (urban or rural), number of family members in the house, and the standard of living index of the household.

**Place of residence:** In general the absence of social interaction of urban living is believed to be associated with the higher risk of violence. In a multi-country study of prevalence and incidence of domestic violence in developing world (Kishor 2004), six out of nine countries, (Cambodia, Columbia, Dominican Republic, Nicaragua, Peru and Zambia) show women living in urban areas are significantly more likely to report domestic violence than rural women. Only two countries (India and Egypt) show opposite relationship.

**Number of people in the household:** Family structure can be considered a contextual setting within which women are empowered to act or are constrained from acting, possibly through the use of domestic violence (Kishor 2000). Previous researches have shown different relationship of perpetuation of domestic violence and the number of people in the household. Some studies indicate that when a woman lives with her in-laws especially in highly patriarchal societies, she is at higher risk of subordination to her husband as well as other members of his family. Some others associate patriarchal extended or joint family living arrangements with less empowerment for women and hence at a higher risk to experience domestic violence. While other studies suggest that women living within a joint family receive a degree of protection from domestic violence given the regular presence of other members of the family in the household (Daga 1998; Visaria 1999; Kishor 2004).

**Standard of living index of the household:** A common assumption in the literature on domestic violence is that women who are poor are more likely to experience violence than women who are not poor (Heise 1998; Jeweks 2002). However, low economic status of the household is not necessarily a causal factor, it is generally assumed to significantly increase the risk of domestic violence. A study at INCLIN (2000) suggests that this variation should be interpreted carefully as women with higher education and from higher income group are less likely to disclose such experiences. Moreover, the direction of the relationship between standard of living and domestic violence is unlikely to be unidirectional. The perpetuation and experience of domestic violence may also contribute to aggravation or even causation of economic instability (Byrne 1999).

Thus with an overview of relationships of various risk factors with the experience of domestic violence, we describe the research design in the following section.

### **Research Design**

The study is divided into two main parts, first, to predict the vulnerability of a woman to experience domestic violence and second, identification of major risk factors associated to the experience of domestic violence incidents among Indian women.

### **Data**

The study is based on the data from India's National Family Health Survey-2 (NFHS-2) conducted in 1998-99. NFHS along with information on fertility, mortality, family planning, and health care, also provides information on violence against women.

The Indian Institute of Population Studies (IIPS) coordinated the survey with financial support from the United States Agency for International Development (USAID) and with additional funding from UNICEF, ORC Macro Calverton at Maryland, USA, and East West Population Center at Hawaii, USA.

NFHS-2 collected information from a nationally representative sample of more than 90,000 ever-married women of age 15-49 years covering India's population over twenty-six states. The analysis is based on ever-married women who responded to the question '*Since you completed 15 years of age, have you been beaten or mistreated physically by any person at home?*' We randomly chose 5,000 observations to fit the predictive models and identify major risk factors for experiencing domestic violence.

### **Variables**

The response variable, "*Has been beaten since age 15 years at home*" is categorical, with "yes" and "no" responses. Based on the conceptual framework, the predictor variables included in the study are women's current marital status, current age, age at first marriage, number of children ever born, education, religion, ethnicity, work status, standard of living index, family structure,

place of residence, husband/partner's education and occupation, spousal age and educational difference and marital duration.

### **Methodology**

We began data analysis using simple cross-tabulations to see which predictor variables are related to instances of domestic violence. However, which risk factors would be most effective would necessarily depend on the consequences associated with the prediction errors. Here there are two kinds of prediction errors, first, failing to predict high-risk women who in reality are at high risk, and second, predicting risk for women who are not at risk. The former can be termed as “false negatives” and the latter can be termed as “false positives”. Thus a risk factor or a group of risk factors that produce few false positives but many false negatives may be discarded if the undesirable consequences from the false negatives are greater than the consequences from the false positives.

As various studies in the literature review suggested, one false negative will cause the same potential harm as several false positives, a number of different “misclassification costs” or ratio of false negatives to false positives were selected for predictive models. We proceeded with five reasonable ratios of the cost of false positives to false negatives, which would cover the range of possible outcome. The ratios were 1 to 1, 2 to 1, 4 to 1, 5 to 1 and 10 to 1.

Using the above mentioned misclassification costs, three predictive models were then developed, namely logistic regression, classification trees (Breiman 1984) and random forest models (Breiman 2001). The predictive accuracy of the models was compared in a classification table using “false negatives” (failing to predict vulnerability for women who really were) and “false positives” (predicting vulnerability for women who were not at risk).

Having described the framework for analysis, the nature of the data and the methodology, we present a detailed analysis of results of the predictive models in the following section.

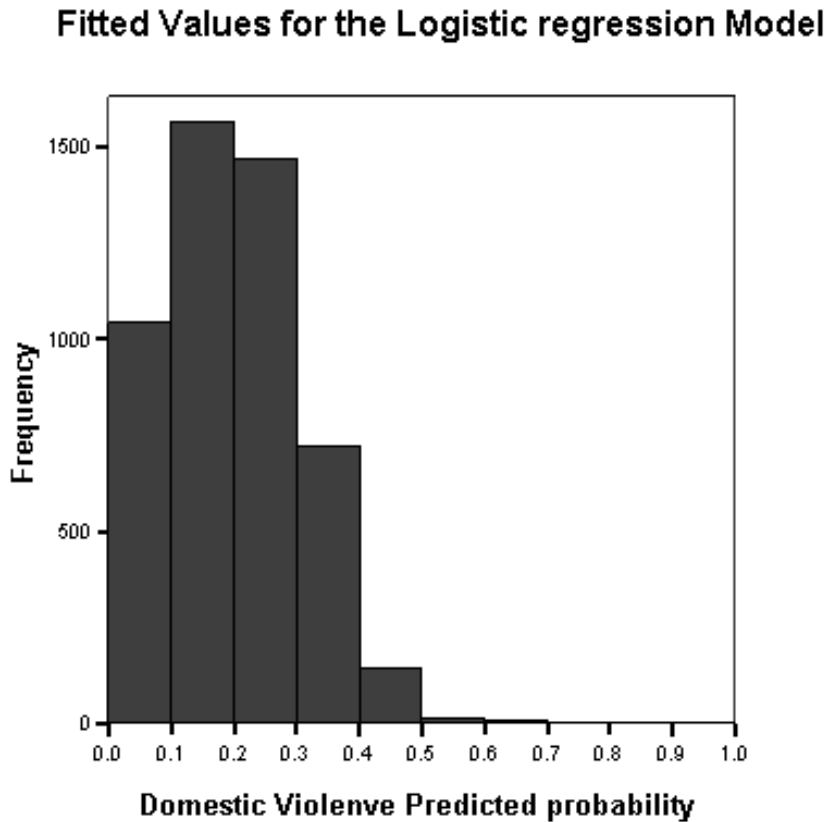
## **Analysis**

### **Building a vulnerability model**

It is common among social scientists to apply logistic regression when the objective is to determine which predictors or risk factors are associated with the binary ('yes' and 'no') outcome of the response variable. However, for this study logistic regression produces the following problems. First, logistic regression results could be used to characterize the data on hand but turning the results into prediction requires additional work. Second, there is no way to effectively incorporate 'misclassification costs' directly into logistic regression despite the fact that they are essential. Third, if one ignores costs and proceeds with logistic regression anyway, the findings could be very misleading.

When we conducted logistic regression model using the most appropriate set of risk factors, only 16 true cases of domestic violence incidents were identified correctly. Figure 3 shows a histogram of the predicted probabilities from the logistic regression model used to classify women into categories of 'yes' and 'no' experience of domestic violence. The graph shows that only a few of these probabilities are greater than 0.5. The 0.5 threshold is important because women with probabilities greater than 0.5 would be classified as vulnerable to experience domestic violence incidents. The most important finding is that only for these few women does the statistical model imply that chances are better than 50-50 of a new woman to experience violence. Thus about 98 percent (970/986) (Table 1) of the true cases are incorrectly determined to have not occurred. Clearly, this is unsatisfactory.

**Figure 3. Probability Distribution of Vulnerability of Women to Domestic Violence from Logistic Regression Model**



**Table 1. Classification Table for the Logistic Regression Model**

**Domestic Violence Experience**

Observed	Predicted as No	Predicted as Yes	Proportion Correct
No	3971	10	0.99
Yes	970	16	0.02

Therefore we turned to new data mining techniques and found that classification tree and random forest performed better than logistic regression at classifying women using the range of misclassification cost ratios mentioned in the methodology section. After running classification

tree and random forest model with all the misclassification cost ratios, the 4 to 1 ratio of the costs of false positives to false negatives produced best results based on a set of predictors that had meaningful association with the experience of domestic violence. The 1 to 1 and 2 to 1 cost ratios generated far too many false negatives, while the 5 to 1 and 10 to 1 cost ratio generated far too many false positives. Figure 4 shows the predictive probabilities of women to experience domestic violence computed by classification tree model using CHAID or Chi-square Adjusted Interaction Detection algorithm. The figure also shows that compared with the earlier results from logistic regression (Figure 3), there are now a substantial number of probabilities greater than 0.5.

**Figure 4. Probability Distribution of Vulnerability of Women to Domestic Violence from Classification Tree Model**

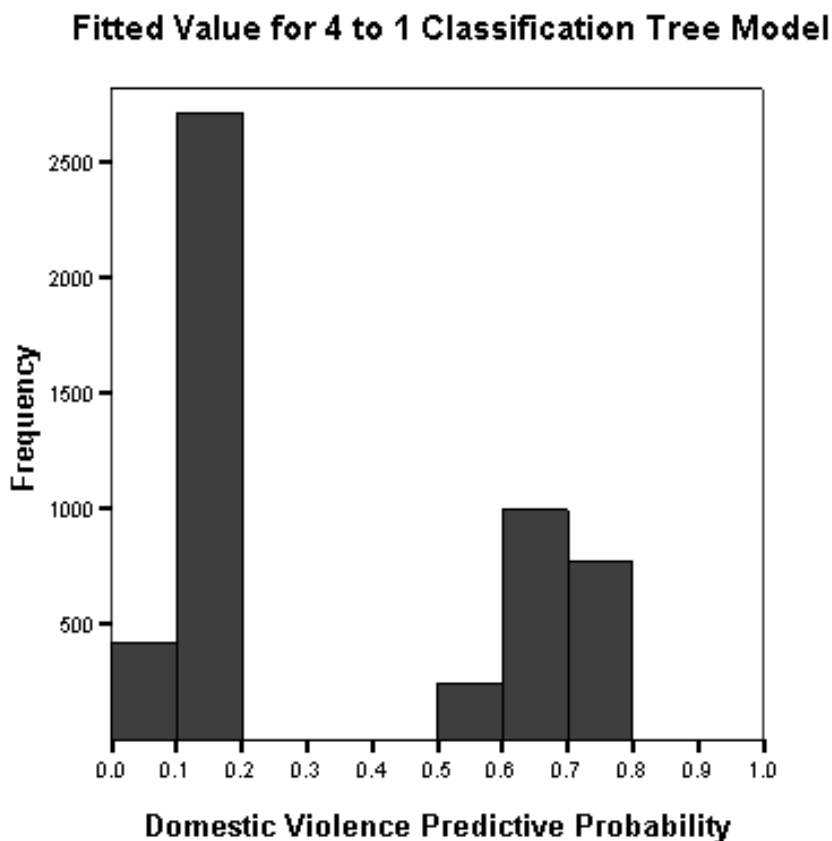


Table 2 is the classification table developed from the results of the tree model. From the first row, we learn that 58 percent of the times women with no experience of domestic violence are correctly identified, and from the second row we could say that for 64 percent of times, women

experiencing domestic violence are correctly identified. These results are a dramatic improvement from the previous logistic model.

**Table 2. Classification Table for the 4 to 1 Classification Tree Model**

**Domestic Violence Experience**

<b>Observed</b>	<b>Predicted as No</b>	<b>Predicted as Yes</b>	<b>Proportion Correct</b>
<b>No</b>	<b>2540</b>	<b>1807</b>	<b>0.58</b>
<b>Yes</b>	<b>398</b>	<b>706</b>	<b>0.64</b>

Another way to interpret the table is to consider the relation between the false negatives and the false positives. There are about 4.5 false positives for every false negative (1807/398), which is virtually close to the 4 to 1 misclassification costs introduced into the model. Because false negatives are 4 times more costly than false positives, the total costs for the two kinds of error balances. This implies that classification tree is performing as specified.

Even though the classification tree results appeared promising, for this kind of data it is sometimes vulnerable to over fitting. More predictions that are accurate could be obtained for a classification tree by using a procedure called “random forests”. The random forest method considers an ensemble of unpruned classification or regression trees created by using bootstrap samples of the training data and random predictor selection at each split (Breiman 2001). A key is that data used to evaluate the strength of the model are not used to build the model. The final prediction for an observation is obtained by aggregating the predictions from the individual trees (often as many as 1000 trees) constituting the ensemble. Given the 4 to 1 misclassification cost ratio, the random forest technique improved the prediction results from the CHAID classification tree (Table 3). This model predicted 61 percent of the women with no experience of domestic violence and 66 percent of the women with experience of violence. Because of the sampling procedure, the total number of cases is slightly different but that does not affect the interpretation.

**Table 3. Classification Table for the 4 to 1 Random Forest Model**

**Domestic Violence Experience**

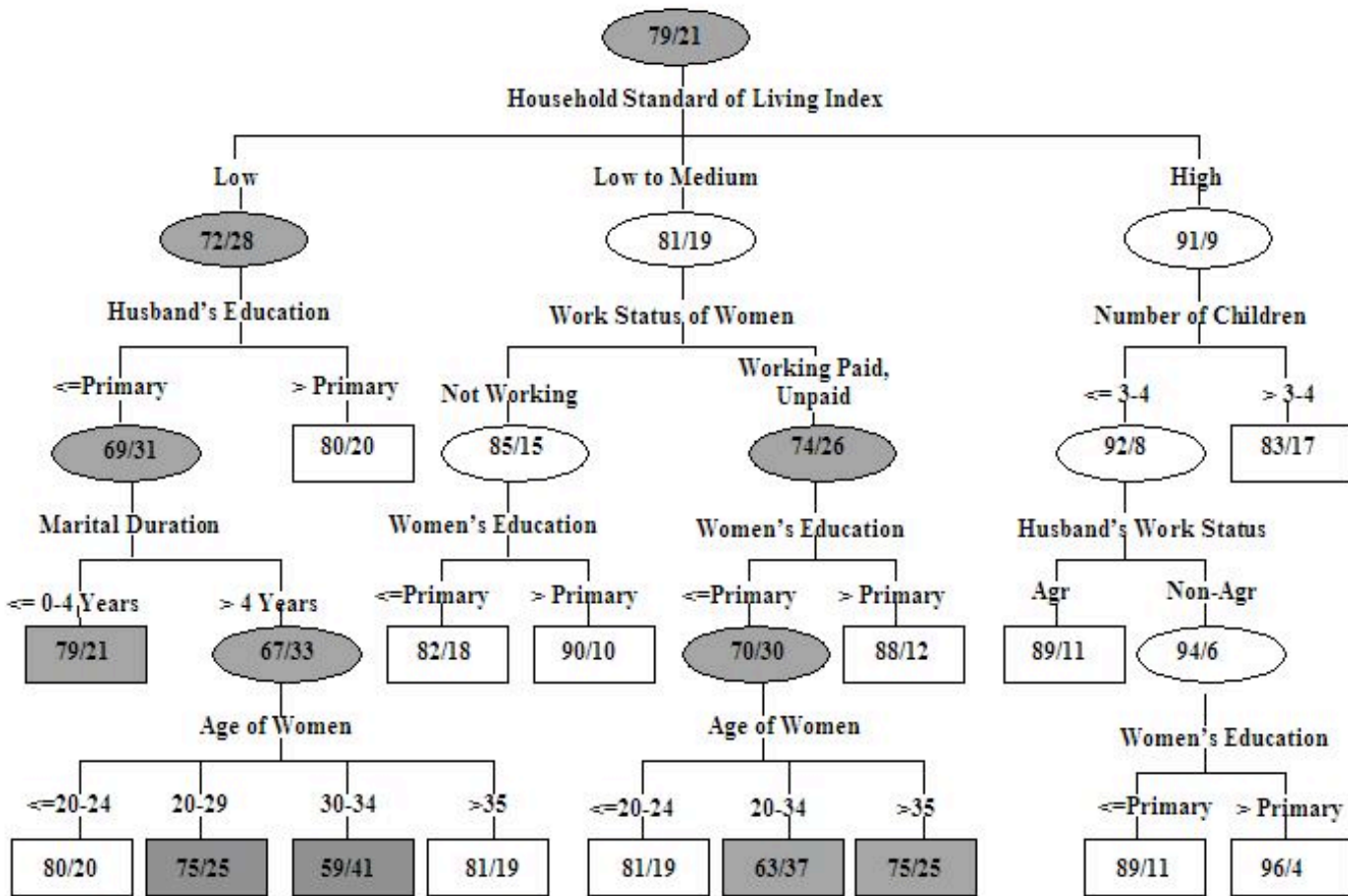
<b>Observed</b>	<b>Predicted as No</b>	<b>Predicted as Yes</b>	<b>Proportion Correct</b>
<b>No</b>	<b>2635</b>	<b>1652</b>	<b>0.61</b>
<b>Yes</b>	<b>372</b>	<b>737</b>	<b>0.66</b>

**Which risk factors are important?**

This section identifies the set of best predictors or risk factors. Figure 5 shows the classification tree produced by CHAID algorithm. At each step, the CHAID chooses a predictor that has the strongest interaction with the response variable. Categories of each predictor variable are merged if they are not significantly different with respect to the response variable. The tree is “read” from top to bottom because that shows the order in which predictors are selected. The ovals represent intermediate subsets of the data while the rectangles represent “terminal” subsets of the data. The color gray means that the “class” of all of the women in that subset or category was classified by the tree algorithm with “positive experience of domestic violence incidents”, while the color white indicates that the “class” of all the women in that group was identified to “negative experience of domestic violence”. The figures in each oval or rectangles show from left to right the actual percentage of women with “no” and “yes” response to the question of experiencing violence.

In figure 5, the topmost oval shows what happens when no predictors are used, there are 79 percent of women without risk of experiencing domestic violence and 21 percent of women with risk of violence. However, given the misclassification cost of 4 to 1, the “parent node” conveys that if no predictors are used, one’s best guess is to classify all women to be vulnerable to domestic violence. As noted earlier, this is the opposite of what we would do if the costs of false negative to false positive were equal. However, prediction results improved further moving down the tree to the “terminal nodes” represented by the rectangles.

**Figure 5. Classification Tree Representation for 4 to 1 CHAID Model**



From figure 5, we can also see that eight risk factors were selected to substantially classify individual women into “yes” and “no” classes of domestic violence experiences. These eight risk factors are: women’s household standard of living index, husband’s education level, marital duration, age of women, women’s status of work, educational level, number of children ever born, and husband’s work status. The tree structure indicates how these risk factors could be used to identify sub groups of women vulnerable to domestic violence in the future. Vulnerability prediction of a subset of women to experience domestic violence could be made under the following three scenarios.

- I. Women belonging to low household standard of living index, and
  1. Husband’s with less than primary education, and
  2. Marital duration less than 4 years.

- II. Women belonging to low household standard of living index, and
  - 1. Husband's with less than primary education, and
  - 2. Marital duration less than 4 years, and
  - 3. Women from 15-29 years of age
- III. Women belonging to low to medium household standard of living index, and
  - 1. Work status of paid and unpaid working, and
  - 2. Women education level less than primary, and
  - 3. Women from 20-34 years of age.

The influence of household standard of living index is not surprising. It makes sense and also there is a common assumption in the past studies on domestic violence that women who are poor are more likely to experience violence than women who are not poor (Heise 1998; Jeweks 2002). The 'low' category of women's standard of living index gives the best split of the data into "yes" and "no" experience of violence among women in our 4 to 1 misclassification cost model. The tree algorithm also considers the complex and non-linear relationship between women's household standard of living index and experience of domestic violence. However, women belonging to low category of household standard of living index could be further subdivided into vulnerable sub groups by other risk factors. The first such risk factors is whether their perpetrators level of education is less than primary. The stakes are raised even higher when marital duration is less than 4 years. However even with marital duration greater than 4 years, and all other previous conditions remaining same, the risk of experiencing domestic violence increases for women of age 15 to 30 years. This is explained by the fact that with increasing age of a women and her marital duration the longer is her period of exposure to the risk of violence.

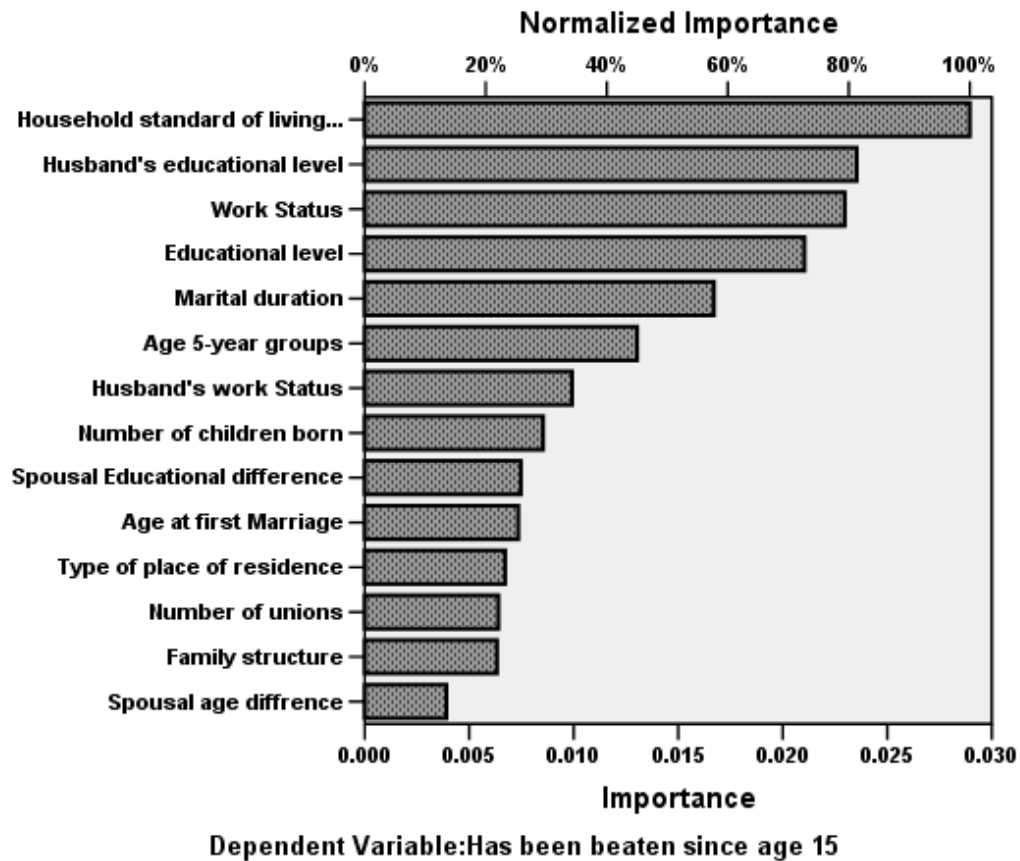
The real power of the analysis is that even when women belonging to 'low to medium' standard of living index is initially identified with no risk of domestic violence, other predictors or risk factors can further determine their vulnerability. From figure 5, we can see that at first level, the 'low to medium' category of women's household standard of living determines the class of women in that intermediate node with no risk of experiencing domestic violence. But if we go down the tree, working status for both paid and unpaid employment of women increases the risk of experiencing violence and then it splits the subset further to classes of "yes" and "no"

experience of domestic violence. The likelihood of risk increases further when women of this particular subset have less than primary level of education and stakes are raised even higher for women of age 20 to 30 years. This is an expected finding. We know from vast literature that education has been one of the sources of empowerment for women. It has given women the ability to gather and assimilate information, manipulate and control the modern world, secure and protect themselves from any form of violence. Thus even with medium standard of living index a woman with low education might have less power, knowledge and skill to protect her from violence. Most interestingly this tree model has conveyed that women engaged in paid and unpaid employment are more likely to be subject to domestic violence than those who are not in the labor force. This could be explained by speculations of various authors that there may be a transition for women who have begun their autonomy. For example, economically active women, may in fact suffer more violence than other women precisely because of the greater control they exert in their own lives, thereby challenging existing patriarchal norms and systems.

It should also be noted that initially we started with 16 possible risk factors but the classification tree model selected only eight important risk factors. No other factors meaningfully improved the results and no other factors improved accuracy if included in addition to the eight factors.

Although, the tree structure gives us an overall picture of the set of important risk factors included in the model, it does not clearly define the importance of each of the selected risk factors or predictors in respect to each other. The random forest model fills this gap. Figure 6 shows the normalized importance of each of the risk factors in predicting the vulnerability of women to experience domestic violence incidents. The list of predictors or risk factors with relatively decreasing importance is as follows: household standard of living index, husband's educational level, work status, educational level of women, marital duration, and age of women.

**Figure 6. Normalized Importance of Risk Factors obtained from Random Forest Model**



### Conclusion and Policy Implications

The conclusions of the paper are as follows. First, using a misclassification cost ratio of 4 to 1 for false positives to false negatives, new data mining models of classification trees and random forest predicted the experience of domestic violence for 66% of women and predicted the absence of such occurrences for 61% of women.

Second, eight risk factors or predictors from a larger pool of risk factors can accomplish this. These eight risk factors are: household standard of living index, husband's education level, marital duration, 5-Year age groups of women, women's status of work, women's educational

level, number of children ever born, and husband's work status. One cannot do meaningfully better by adding more predictors or risk factors.

Third, we could precisely identify vulnerable target groups and sub-groups of women. For example, even when group of women belonging to "low to medium" standard of living index were classified with no risk of experiencing domestic violence incidents by the tree at the first split, the following splits made by meaningful predictors identified vulnerable subgroups of women. This subgroup of women has the following characteristics, paid and unpaid working status, less than primary level of education, and belonging to 20 to 30 years of age. Here lies the advantage of using classification trees.

Fourth, the extra information of importance of variables from random forest model could lead to further research on causal factors of domestic violence.

Fifth, we believe that the models developed in this study can serve as tools to support the development of quick-response programs or even preventive measures for policy makers, social workers, NGO's, law enforcement assistance and other researchers to identify groups and sub groups of women to experience domestic violence incidents.

However, we should also consider some important caveats of this study. First, under reporting of cases of domestic violence due to social indignation, lack of awareness, and knowledge can affect the prediction results. Second, although, it may be tempting to infer that our predictors and risk factors are important causes of domestic violence, we counsel great caution. For example, the role of economic status of women in experiencing violence among is not clear.

Finally, prediction is necessarily data dependent. It is very likely that prediction strength and accuracies would differ with different data sets, nature of risk factors and socio-economic and cultural settings of different regions.

### References

- Bhatti, R.S. 1990. Socio-cultural Dynamics of Wife Battery. In *Violence Against Women*, edited by S. Sood. Jaipur, India: Arahant Publishers.
- Breiman, L. 2001. Random Forests. *Machine Learning* 45:5-32.
- Breiman, L., Friedman, J.H., Olshen, R.A., and C.J. Stone. 1984. *Classification and Regression Trees*. Monterey, Ca: Wadsworth.
- Byrne, C.A., H.S. Resnick, D.G. Kilpatrick, C.L. Best, and B.E. Saunders. 1999. The socioeconomic impact of interpersonal violence on women. *Journal of Consulting and Clinical Psychology* 67 (3):362-366.
- Daga, A.S., S. J. Jejeebhoy, et al. 1998. Preventing violence, caring for survivors: Role of health professionals and services in violence. In *Domestic Violence against women: An investigation of hospital casualty records*. Mumbai, India: Research Center of Ausandhan Trust.
- Dyson, T., and M. Moore. 1983. On kinshipstructure, female autonomy and demographic behavior in India. *Population and Development Review* 9 (1):35-60.
- Ellsberg, M.C. 2000. *Candies in hell: Research and action on domestic violence against women in Nicaragua*. Umea, Sweden: Umea Univesity.
- Garcia, B., ed. 2000. *Women, poverty and demographic change*. Liege: Oxford University Press for IUSSP.
- Heise, L.L. 1998. Violence Against Women: An integrated, ecological framework. *Violence Against Women* 4 (3):262-290.
- Hornung, C.A., B.C. McCullough, and T. Sugimoto. 1981. Status relationships in marriage: Risk factors in spouse abuse. *Journal of Marriage and the Family* 43:675-692.
- IIPS, and Macro. 2000. National family Health Survey (NFHS-2), 1998-99: India. Mumbai: International Institute for Population Sciences.
- INCLN. 2000. Domestic Violence in India 3: A Summary Report of a Multi-Site Household Survey. Washington, DC.: Intrenational Center for Research on Women and The Center for Development and Population Activities.
- Jejeebhoy, Shireen. 1998. Wife-beating in rural India: A husband's right? *Economic and Political Weekly* 33 (15):855-862.

- Jeweiks, R. 2002. Intimate partner violence: Causes and prevention. *Lancet* 359 (9315):1423-1429.
- Johnson, K. 2003. Dialectics of power and violence in the home: A comparative analysis of women's experience of domestic violence in Haiti and Nicaragua., Dissertation. University of Maryland.
- Kishor, S. 2000. Empowerment of women in Egypt and links to the survival and health of their infants. In *Women's empowerment and demographic processes: Moving beyond Cairo.*, edited by G. Sen, and H.B. Presser. New York: Oxford University Press.
- Kishor, S., and Johnson, K. 2004. *Profiling Domestic Violence - A Multi-Country Study*. Calverton, Maryland: ORC Macro.
- Malhotra, A., and M. Mather. 1997. Do schooling and work empower women in developing countries? The case of Sri Lanka. *Sociological Forum* 12 (4):599-630.
- Martin, S. L., A.O Tsui, K. Maitra, and Marinshaw. 1999. Domestic Violence in northern India. *American Journal of Epidemiology* 150 (4):417-426.
- Mason, K.O. 1987. The impact of women's social position on fertility indeveloping countries. *Sociological Forum* 2 (4):718-745.
- Miller, B.D. 1999. Wife Beating in India: Variations on a Theme. In *To have and to hit*, edited by a. J. C. C. J.K.B. Dorothy Ayers Counts. Urbana and Chicago: University of Illinois Press.
- Mitra, N. 1999. Best Practices among Responses to Domestic Violence in Maharashtra and Madhya Pradesh. Domestic Violence in India: A Summary Report of Three Studies. Washington, DC.: International Center for Research on Women.
- Mueller, C.W., T.L. Parcel, and F.C. Pampel. 1979. The effect of marital dyad status inconsistency on women's support for equal rights. *Journal of Marriage and the Family* 56:1121-1139.
- NCRB. 1999. National Crime Records Bureau, 1995-99. Delhi: National Crime Records Bureau of India.
- Rao, S., I. S., et al. 2000. Domestic Violence: A Study of Organizational Data. Domestic Violence in India: A Summary Report of Four Records Studies. Washington, DC: International Center for Research on Women.

- Verma, R.K. and M. Collumbien. 2003. Wife beating and the link with poor seual health and risk behavior among men in urban slums in India. *Journal of Comparative Family Studies* 34 (1):61-74.
- Vindhya, U. 2000. Dowry deaths in Andhra Pradesh, India: response of the criminal justice system. *Violence Against Women* 6 (10):1085-1108.
- Visaria, L. 1999. Violence against Women in India : Evidence from Rural Gujarat. Domestic Violence in India: A Summary Report of Three Studies. Washington, DC.: International Center for Research of Women.