January 2005

Developing a Practical Forecasting Screener for Domestic Violence Incidents

Richard A. Berk
University of Pennsylvania, berkr@sas.upenn.edu

Yan He
University of California

Susan B. Sorenson
University of Pennsylvania, sorenson@sp2.upenn.edu

© 2005 Sage Publications. Postprint version. Published in *Evaluation Review*, Volume 29, Issue 4, pages 358-383 Publisher URL: 10.1177/0193841X05275333.

NOTE: At the time of publication, authors Susan B. Sorenson and Richard A. Berk were affiliated with the University of California. Currently (August 2007), she is a faculty member in the School of Social Policy and Practice, and he is a faculty member in the Department of Criminology at the University of Pennsylvania.

Developing a Practical Forecasting Screener for Domestic Violence Incidents

Abstract

In this paper, we report on the development of a short screening tool that deputies in the Los Angeles Sheriff's Department could use in the field to help forecast domestic violence incidents in particular households. The data come from over 500 households to which sheriff's deputies were dispatched in the fall of 2003. Information on potential predictors was collected at the scene. Outcomes were measured during a three month follow-up. The data were analyzed with modern data mining procedures in which true forecasts were evaluated. A screening instrument was then developed based on a small fraction of the information collected. Making the screening instrument more complicated did not improve forecasting skill. Taking the relative costs of false positives and false negatives into account, the instrument correctly forecasted future calls for service about 60% of the time. Future calls involving domestic violence misdemeanors and felonies were correctly forecast about 50% of the time. The 50% figure is especially important because such calls require a law enforcement response and yet are a relatively small fraction of all domestic violence calls for service.

A number of broader policy implications follow. It is feasible to construct a quick-response, domestic violence screener that is practical to deploy and that can forecast with useful skill. More informed decisions by police officers in the field can follow. Although the same kinds of predictors are likely to be effective in a wide variety of jurisdictions, the particular indicators selected will vary in response to local demographics and the local costs of forecasting errors. It is also feasible to evaluate such quick-response threat assessment tools for their forecasting accuracy. But, the costs of forecasting errors must be taken into account. Also, when the data used to build the forecasting instrument are also used to evaluate its accuracy, inflated estimates of forecasting skill are likely.

Keywords

domestic violence, data mining, police, spousal assault, family violence

NOTE: At the time of publication, authors Susan B. Sorenson and Richard A. Berk were affiliated with the University of California. Currently (August 2007), she is a faculty member in the School of Social Policy and Practice, and he is a faculty member in the Department of Criminology at the University of Pennsylvania.

Developing a Practical Forecasting Screener for Domestic Violence Incidents*

Richard Berk and Yan He
Department of Statistics, UCLA

Susan B. Sorenson
School of Public Health, UCLA

January 25, 2005

Abstract

In this paper, we report on the development of a short screening tool that deputies in the Los Angeles Sheriff's Department could use in the field to help forecast domestic violence incidents in particular households. The data come from over 500 households to which sheriff's deputies were dispatched in the fall of 2003. Information on potential predictors was collected at the scene. Outcomes were measured during a three month follow-up. The data were analyzed with modern data mining procedures in which true forecasts were evaluated. A screening instrument was then developed based on a small fraction of the information collected. Making the screening instrument more complicated did not improve forecasting skill. Taking the relative costs of false positives and false negatives into account, the instrument correctly forecasted future calls for service about 60% of

*This project would have been impossible to undertake without the hard work of Sergeant Robert Jonsen, Deputy Cecilia Ramirez, Lieutenant Charles Stringham, and Sergeant Christopher Cale, all of the Los Angeles Sheriff's Department. Thanks also go to the deputies who helped in the data collection. The paper is heavily based on a report written for the Los Angeles Sheriff's Department, available as preprint #390 from the UCLA Department of Statistics home page (www.stat.ucla.edu) under "publications."

the time. Future calls involving domestic violence misdemeanors and felonies were correctly forecast about 50% of the time. The 50% figure is especially important because such calls require a law enforcement response and yet are a relatively small fraction of all domestic violence calls for service.

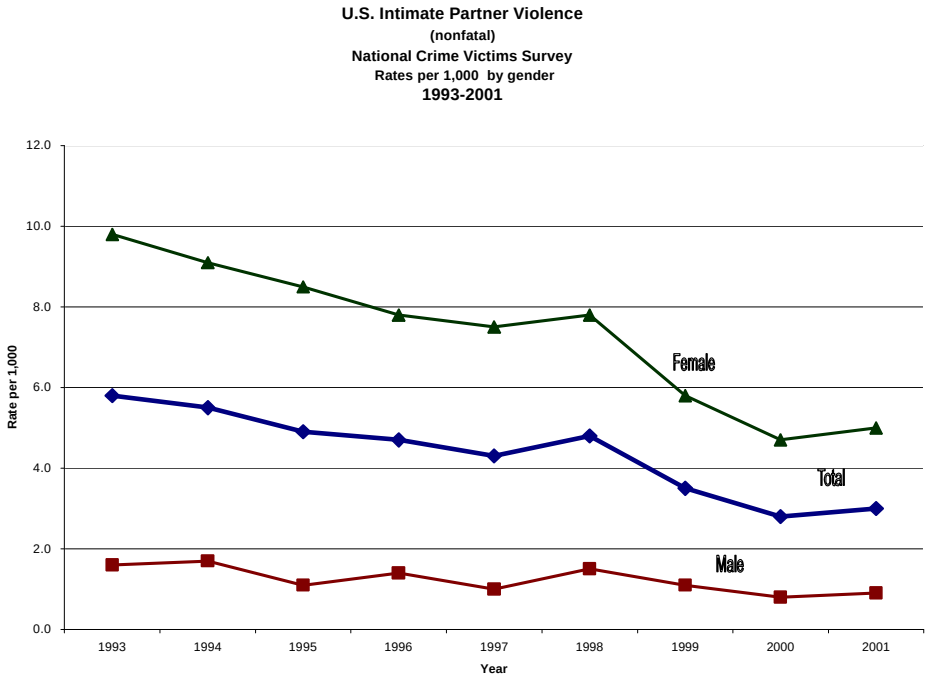
A number of broader policy implications follow. It is feasible to construct a quick-response, domestic violence screener that is practical to deploy and that can forecast with useful skill. More informed decisions by police officers in the field can follow. Although the same kinds of predictors are likely to be effective in a wide variety of jurisdictions, the particular indicators selected will vary in response to local demographics and the local costs of forecasting errors. It is also feasible to evaluate such quick-response threat assessment tools for their forecasting accuracy. But, the costs of forecasting errors must be taken into account. Also, when the data used to build the forecasting instrument are also used to evaluate its accuracy, inflated estimates of forecasting skill are likely.

1 Introduction

Domestic violence remains a very serious public health and law enforcement problem, especially for women. For the country as a whole, Figure 1 shows that the general declines in nonfatal domestic violence since the early 1990's have recently leveled off and that women are still far more likely than men to be victims. Figure 2 shows much the same pattern for partner homicides in California. Although law enforcement is not called to most domestic violence scenes, many law enforcement calls involve domestic violence. For example, in California in 2002, there were 196,569 domestic violence calls to law enforcement, 119,850 involving weapons, which also represents a leveling off since the declines of the 1990s (California Department of Justice, 2003). From the enormous number of calls for service alone, it follows that law enforcement resources for domestic violence incidents should be allocated in an efficient manner.

When responding to a domestic violence incident, police officers can choose from a long menu of options such as making an arrest, ordering the offender from the premises, referring the offender to a treatment program for batterers, transporting the victim to a shelter, or just trying to restore order (Sherman, 1992). But the most useful response depends in part on whether

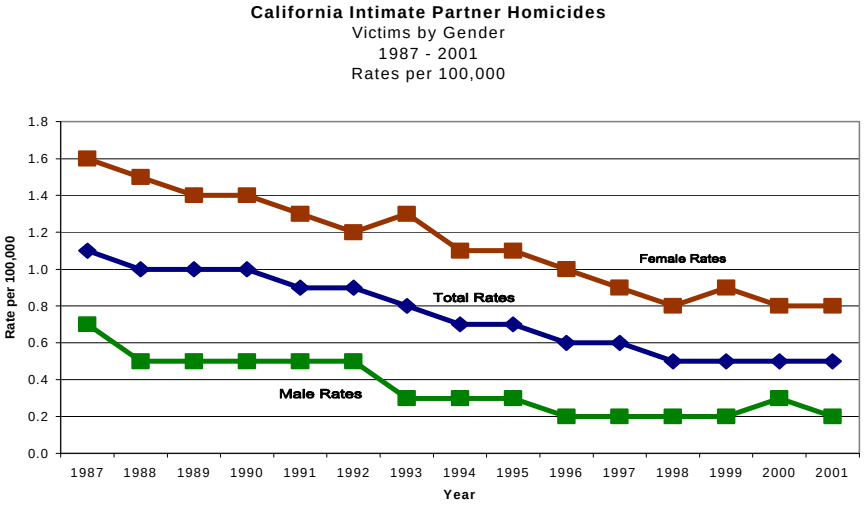
new calls for service are likely. Quite properly, an incident seen as a one-time event often will be treated differently from an incident that is likely to be repeated. Just as accurate forecasts of crime "hot spots" can be useful when decisions are made about where to patrol (Skogan and Frydl, 2002), accurate forecasts of domestic violence calls for particular households can assist police officers at the scene to decide better what action to take.



	1993	1994	1995	1996	1997	1998	1999	2000	2001
Total	5.8	5.5	4.9	4.7	4.3	4.8	3.5	2.8	3.0
Male	1.6	1.7	1.1	1.4	1.0	1.5	1.1	0.8	0.9
Female	9.8	9.1	8.5	7.8	7.5	7.8	5.8	4.7	5.0

Figure 1: Intimate Partner Violence in the US

In this paper, we consider a project undertaken for the Los Angeles County Sheriff's Department in which the goal was to develop a short screen-



	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Total Rates	1.1	1.0	1.0	1.0	0.9	0.9	0.8	0.7	0.7	0.6	0.6	0.5	0.5	0.5	0.5
Male victims	0.7	0.5	0.5	0.5	0.5	0.5	0.3	0.3	0.3	0.2	0.2	0.2	0.2	0.3	0.2
Female victims	1.6	1.5	1.4	1.4	1.3	1.2	1.3	1.1	1.1	1.0	0.9	0.8	0.9	0.8	0.8

Source: California Dept. of Justice, Criminal Justice Statistics Center, Homicides 1987-2001, ages 20-44, July 2003.

Figure 2: Intimate Partner Homicides in California

ing instrument that would help sheriff's deputies in the field predict future domestic violence incidents and their seriousness. We did not intend to develop a state-of-the-art forecasting tool to predict such things as "lethality," nor a monitoring protocol for long term case management. A device of either sort would have required far more resources and would have been impractical to employ in the field. We hoped to find a set of approximately five questions that deputies could administer and score very quickly at the scene and that could help them anticipate better whether future domestic violence was likely.

2 Past Research

"Threat assessment tools" for domestic violence have been under development for at least a decade (Fein et al., 1995). Some are currently in use. Among the more visible examples of domestic violence threat assessment instruments that are now in the field are the following:

1. Mosaic Threat Assessment System (Galvin de Becker & Associates, Los Angeles, California);

2. Conflict Tactics Scale (Military Family Resources Center, Arlington, Virginia)

3. Kingston Screening Instrument (The Connecticut Office of Policy and Management, Hartford, Connecticut);

4. Lethality Checklist and Physical Abuse Scale (Probation Department, Monterey County, California);

5. Domestic Violence Referral Form (Domestic Violence Enhanced Response Team, Colorado Springs, Colorado);

6. Pre-Sentencing Investigation DV Supplement (Duluth, Minnesota);

7. Risk Assessment and Lethality Assessment (First Judicial Circuit, Hawaii); and

8. Spousal Assault Risk Assessment Guide (British Columbia Institute on Family Violence, British Columbia, Canada).

The items included in these instruments are typically consistent with past research on domestic violence (Campbell et al., 2003; Carraneo and Goodman, 2003) and research on domestic violence calls to the police (Berk et al., 1984; Houry et al., 2004). They are also significantly grounded in the relevant social science theory (e.g., Straus and Gelles, 1990; Sherman et al., 1992). For example, the degree to which the batterer has a stake in "conformity" is commonly measured. Finally, the items are, by and large, consistent with the experience of seasoned police officers. Taken as a whole, however, the existing instruments have several significant weaknesses as quick-response forecasting tools.

First, we required a very short screener that sheriff's deputies could use in the field. All of the available instruments we could find were far too long and complex to apply at the scene of a domestic violence incident. There was too much information being sought and no quick way to arrive at an accurate overall assessment.

Second, our goal was forecasting only. We were not seeking to understand the causes of the violence. Nor were we concerned about issues beyond law enforcement considerations. So, many of the commonly included items were irrelevant.

Third, we needed a forecasting screener that had been developed and evaluated taking into account the very heterogeneous mix of ethnic groups and recent immigrants served by the Los Angeles Sheriff's Department. Anglo and Latino households are the most common, but there are significant numbers of African-American households as well as households from many different countries in Asia (e.g., Russia, Armenia, China, Korea, Vietnam, Philippines, Cambodia, and Laos). An instrument developed, for example, for white and African-American households only might forecast less accurately than one hand-tailored for the Los Angeles area.

Fourth, some of the available instruments were not designed as forecasting tools. For example, a spokesperson for Mosaic, when asked by e-mail whether forecasting was a goal, responded *"For the record, MOSAIC is NOT a predictive instrument. It is a tool designed to assist those making assessments and case management decisions in gathering relevant information and putting that information in the context of the situation"*[1] (personal communication from Robert Martin at info@mosaicsystem.com, emphasis in the

[1]The e-mail note goes on to say, *"To date, there are no published attempts to 'validate' MOSAIC as it is not a model that lends itself well to statistical methods - simply because it is not statistically based....I can say that in a non-statistical arena, MOSAIC is 'validated'*

original e-mail). For most other instruments, the importance of forecasting was unclear.

Finally, we could find no evaluations of the available instruments that properly addressed forecasting skill (Dutton and Kropp, 2000; Roehl and Guertin, 2000). An appropriate test of how well an instrument forecasts requires that the data used to assess forecasting ability not be the data used build the statistical model. It is well known that if the same data are used for both purposes, "internal" performance measures will likely produce a falsely optimistic evaluation of how well the model forecasts (Efron and Tibshirani, 1993: chapter 17). Thus, even if the outcomes are measured later in time than the predictors, conventional measures of fit, model diagnostics, or statistical tests will not accurately characterize forecasting skill. Moreover, for forecasts to be useful for making real world decisions, the relative costs of false positives and false negatives need to be built into the forecasting enterprise. To ignore the costs of false positives and false negatives is to proceed as if the costs were the same. Rarely do equal costs make much policy sense and typically, when the costs of forecasting errors change, so do the forecasts.[2]

In summary, the research we undertook for the Los Angeles Sheriff's Department was in response to their need for a short domestic violence screener that could be used to make forecasts in the field. Although we could draw on insights from the existing literature, we could find no "off-the-shelf" instrument that demonstrably could do the job.

3 Research Design

The research design specified a representative sample of 1500 households. These were to be households to which sheriff's deputies had been dispatched for incidents that were likely to involve domestic violence. The deputies were to employ a screener of about 30 questions (see Appendix A) as part

hundreds of times a day by practitioners managing cases. While not a formal study, I can tell you 100% of the people who use MOSAIC that we questioned informed us that by using the MOSAIC, their assessments, and subsequently their case management decisions, are better informed than without it."

[2]Jacquelyn C. Campbell and several colleagues are currently engaged in a project, funded by the National Institute of Justice, titled "Intimate Partner Violence Risk Assessment Instruments: A Prospective Validation Field Experiment." It is not clear whether actual forecasting skill is being evaluated and whether the costs of forecasting errors are being taken into account. Results from the study are not available at this time.

of their usual duties at the scene. The questions to be asked were selected because of their perceived importance in past research and their potential to predict domestic violence in the future. For example, researchers and police officers speak with one voice about the importance of past domestic violence incidents as a good predictor of future domestic violence incidents. We anticipated that in the field, the victim would be the primary source of information but that information would be obtained from others as well.

In a three-month follow-up period, all new dispatches to the 1500 household were to be recorded. The answers to the screener would then, with the help of new data mining procedures, be used to forecast which households had subsequent calls and how serious those incidents were. It was anticipated that a few of the items would prove to be far better predictors than others. Those better predictors would then comprise the short screener recommended to the Los Angeles Sheriff's department.

Three features of the data need to be emphasized. First, the outcome of interest was a call to the same household during the three-month follow-up period. While these calls were likely to be for a domestic dispute, the calls could involve other law enforcement concerns. The boundaries between domestic violence and other household incidents are sometimes unclear, and we wanted initially to cast a wide net.

Second, the data collected with the screener was based largely on perceptions of victims and others at the scene. For the screener to be employed in the field, it had to be based on information that was readily available and that could be obtained within usual law enforcement practice. Whether the perceptions of the victim and others were always fully accurate does not matter as long as useful forecasts follow. In the same spirit, the screener items were not collected to unravel the many causes of domestic violence. That would entail a different study.

Third, the study was mounted in six substations selected, as a matter of efficiency, because they accounted for largest numbers of domestic violence calls.[3] It was assumed that all deputies in these substations would cooperate. However, it took far longer than expected for the project staff within the Sheriff's Department to obtain permission to field the screener. Subsequent cooperation from deputies was spotty. A key reason was that the study was begun in the midst of a work slowdown motivated by a contract dispute with

[3]The substations were Century City, Compton, East Los Angeles, City of Industry, Lakewood, and Lancaster.

Los Angeles County. In the end, deputies collected on-the-scene data from fewer than half the number of households specified by the research design. It is clear that deputies were able to exercise wide discretion in when to use the screeners, but we do not know what consequences there are, if any, for how representative the sample of households is. We turn to some tabulations now that may help to address this issue.

4 Data from the Long Screener

The tabulation from the long screener provides important background information on the kinds of domestic violence cases to which sheriff's deputies are dispatched. The information also provides a context in which the forecasting can be undertaken.

Nearly three-quarters of the households in this study had experienced domestic violence in the past. For these households, the most recent prior occurrence was within the preceding 6 months. For about half of these households, the police had been called twice before or more. According to what the deputies could learn at the scene, about a quarter of the time an earlier incident had led to an arrest and about 15% of the time, a conviction for domestic violence followed. For nearly three-quarters of these households, the violence is reported to be getting worse with nearly a quarter of the victims seeking medical attention from a previous assault. About 16% said they were treated in an emergency room.

About half of the perpetrators are reported to keep track of whom the victim talks to on the phone and/or to determine which friends the victim can see. In addition, a solid majority are reported to have problems with jealousy, drugs and drinking. About 60% destroy property in the household when angry and about a third have threatened to kill someone in the victim's family. In about 10% of the households with children, one or more of the children are reported to have been injured by the perpetrator when he or she is angry. About 10% of the perpetrators are said to have a handgun and a bit more than 40% of those are said to have threatened the victim with it in the past. About 4% of the perpetrators are reported to have a rifle and a bit less than 40% are said to have threatened the victim with it in the past.

Restraining orders were reported to be in place in approximately 10% of the households. A little more than half of the victims said that they left the perpetrator in the past. Among those who had left in the past, about half

had left 2 or more times. A bit less than half of the perpetrators are reported to have regular jobs.

These tabulations are about what one would expect from a sample of households to which deputies are called to resolve a domestic dispute. The figures imply that generalizations can be usefully made from the sample to all such households served by the Los Angeles Sheriff's Department. But, there are no guarantees, and the results to follow must be treated with a bit more caution than had the research design been implemented fully as intended.

5 Forecasting Calls for Service

5.1 Missing Data

We had data from the screening instrument for 671 households. However, some of the items on that instrument were not filled out, and if households with missing data were eliminated (i.e., listwise deletion), there would be 516 complete observations.

We applied three approaches to the missing data. First we experimented with imputation of the missing data before the statistical analyses began. However, imputation with data mining is quite new and not yet fully accepted. Moreover, the procedures we used produced results that were quite unstable across bootstrap samples of the data. We decided that the cure was worse than the disease.

Second, we coded the missing data with their own indicator variables. Missing data became a legitimate category for each predictor with incomplete data. Because for most variables the amount of missing data was small, this too produced unstable results; the variance for each missing data indicator variable was small. In addition, the results were difficult to interpret. Again the cure was worse than the disease.

Finally, we decided that listwise deletion was the most prudent course, especially because we lost only about 15% of the households. With listwise deletion we arrived at results that were stable, made sense, and forecasted remarkably well.

For the 516 households with complete data, there was at least one return call for 109. Thus, about 21% had a return call within three months after the screener information was collected. Although there is no guarantee that all of these calls were for domestic violence incidents, no doubt most were.

5.2 Taking Costs into Account

We began the analysis using simple cross-tabulations to see which screener items were related to whether there were subsequent calls for the households studied. It was apparent that some items had considerable promise. However, which items would be most effective would necessarily depend on the consequences associated with forecasting errors. In this instance, there were two kinds: 1) failing to predict high risk for households that really were and 2) predicting high risk for households that really were not at high risk. The former can be viewed as "false negatives" and the latter can be viewed as "false positives." Thus, a predictor that produced few false positives but many false negatives might be discarded if the undesirable consequences from the false negatives were larger than the undesirable consequences from the false positives. We needed, therefore, information from the Los Angeles Sheriff's Department on the consequences of false positives and false negatives.

Efforts to elicit this information from the Los Angles Sheriff's Department led to a general conclusion that false negatives were substantially more problematic than false positives. In other words, they considered not responding to a call when there actually was a need for law enforcement assistance more of a problem than rolling on a call that turned out to be, in essence, a false alarm. One false negative would produce about the same potential harm as several false positives, but the precise figures for these "costs" could not be determined. As a technical matter, all we needed for our statistical procedures was the ratio of false negative costs to false positive costs, but this was also too demanding. We proceeded, therefore, with four reasonable, but different, ratios of the costs of false negatives to the costs of false positives that would cover the range of likely values: 1 to 1, 2 to 1, 5 to 1, and 10 to 1. Consistent with the information provided by the Sheriff's Department, for none of the ratios was the failure to accurately forecast a new call for service more costly than incorrectly forecasting a new call for service.

One can gain further understanding about the key role of costs using the obtained 21% return call figure. If for every household, one predicted another call within three months, one would be correct about 21% of the time. And, one would also be wrong about 79% of the time. Conversely, if for every household, one predicted no calls within three months, one would be correct about 79% of the time. And one would also be wrong about 21% of the time. Which is a better strategy: always predicting a future call or not? The answer depends on the costs of false negatives compared to the costs of false

positives.

If both were equally costly, the best strategy would clearly be to never predict a subsequent call. But now suppose that failing to anticipate future calls was very costly; suppose that these false negatives were 10 times more costly than false positives. Then, the best strategy would clearly be to always predict a subsequent call.[4] In short, the relative costs of false negatives compared to the relative costs of false positives can affect how forecasting is done. And it also affects, therefore, which predictors are likely to be important.

5.3 Building a Forecasting Model

Credible forecasting of calls for service requires two steps. First, strong associations are required between information contained in the screener and whether there were subsequent calls during the three-month follow-up period. Finding these associations is a task for multivariate statistics. Second, once some strong associations are found, these can be used in the future to link screener information from new households to the chances of return calls to these new households. The assumption is that the associations found in the data on hand apply to new households in the future. This is just another way of saying that in all forecasting, whatever the application, success depends on the future being substantially like the past. For example, if the mix of households served by the Sheriff's Department changed dramatically, forecasts based on procedures developed with the current data would be suspect.

Focusing initially on the first step, it is common for criminologists to apply logistic regression when the goal is to determine which predictors are associated with outcome such as ours (e.g., Campbell et al., 2003). For our enterprise, however, logistic regression produces four problems. First, logistic regression can be used to characterize the data on hand, but turning the results into forecasts requires additional work. Second, there is no way to effectively introduce 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. One has implicitly assumed that the costs of false negatives and false positives are the same.

[4]This is because $(10 \times .21) > (1 \times .79)$. When the costs are the same $(1 \times .21) < (1 \times .79)$. One can think of these calculations as providing comparisons between the expected costs of different forecasting strategies.

Finally, as an empirical matter, logistic regression does not help much with these data.

When we applied logistic regression using the most promising predictors available, only 10 true subsequent calls out of 109 were identified correctly as such. Figure 3 shows a histogram of the probabilities from the logistic regression used to identify households with one or more calls for service during the follow-up period. One can see that only a few of these probabilities are larger than .50. The .50 threshold is important because households with probabilities greater than .50 would be classified as having subsequent calls. For these households the chances are better than 50-50. The key message is that only for these very few households does the statistical model imply that the chances are better than 50-50 of a future call for service. Thus, about 91% (99/109) of the true calls are incorrectly determined to have not occurred. Clearly, this is unsatisfactory.

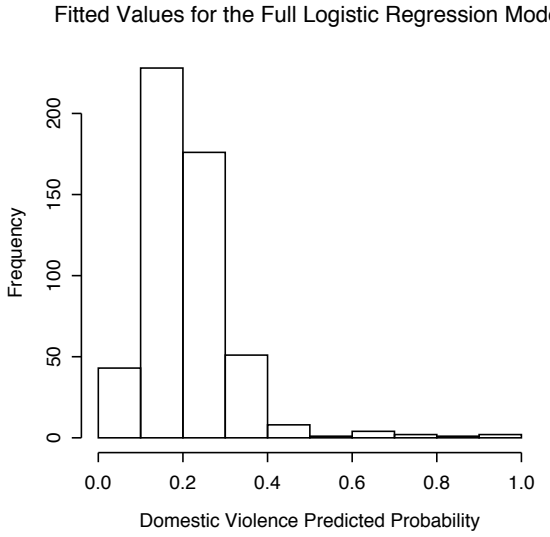


Figure 3: Probability Distribution of A New Call Produced by Logistic Regression

Therefore, we turned to data mining techniques and found that Classification and Regression Trees (CART) performed far better than logistic regression at classifying households, using the range of cost ratios elicited

from the Sheriff's Department. The 5 to 1 ratio of the costs of false negatives to false positives produced results based on a sensible set of predictors that had useful associations with the calls for service during the follow-up period. The 1 to 1 and 2 to 1 cost ratios generated far too many false negatives, while the 10 to 1 cost ratio generated far too many false positives. Figure 4 shows that, compared to the earlier results for logistic regression, there are now a substantial number of probabilities greater than .50.

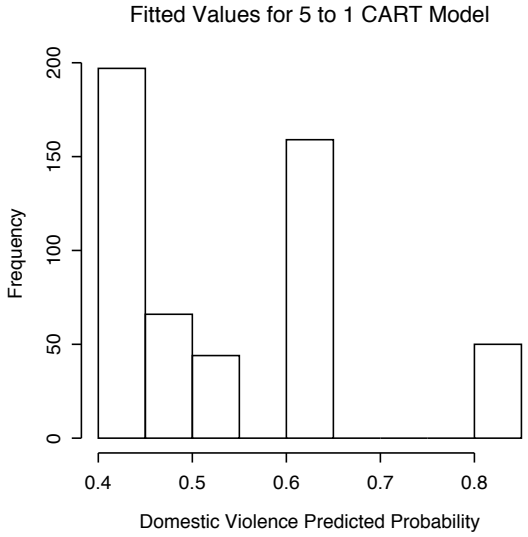


Figure 4: Probability Distribution for New Call Produced by the 5-to-1 CART Model

How well do we now sort households into those that had new calls and those that did not? Table 1 shows the relevant "classification table." From the first row we learn that 56% of the time households with no subsequent calls are correctly identified. From the second row we learn that 66% of the time households that have a subsequent call are correctly identified. These results are a dramatic improvement.

Another way to look at the table is to consider the relationship between the false negatives and the false positives. There are about 4.9 false positives for every false negative (181/37), which is virtually the same as the 5 to 1 costs introduced into the model. Because false negatives are taken to be 5

times more costly than false positives, the *total* costs for the two kinds of errors balance. This implies that CART is performing as intended.

	Identified as No Call	Identified as Call	Proportion Correct
No Call	226	181	0.56
Call	37	72	0.66

Table 1: Classification Table for the 5-to-1 CART Model

5.4 Which Screener Items Work?

Figure 5 shows the classification tree produced by CART. The tree is "read" from top to bottom because that reflects the order in which screener items were selected. The ovals represent intermediate subsets of the data while the rectangles represent "terminal" subsets of the data. The letter "b" means that all of the households in that group were identified by CART as "call households" while the letter "a" means that all households in that group were identified by CART as "no call households." The figures in each oval or rectangle show from left to right the actual number of households without a subsequent call and the actual number of households with a subsequent call. At the top, for example, which shows what happens when no predictors are used, there are 410 households without a subsequent call and 109 household with a subsequent call. Given the 5 to 1 cost differential, that "node" conveys that if no predictors are used, one's best guess is to identify all households as "call households." As noted earlier, this is the opposite of what one would do if the costs of false negatives and false positives were equal. But one can do a lot better moving down the tree to the terminal nodes represented by the rectangles.

From Figure 5, one can see that four screener items were selected because they substantially help sorting households into "call" and "no call" groups. These four items, in the order selected are: the reported number of previous calls to that household, whether the perpetrator was reported to destroy household property when angry, whether the perpetrator was reported to be unemployed, and whether the perpetrator was reported to have threatened to kill the victim or someone else in the family in the past. The tree structure conveys how these items can be used to forecast calls for service from new

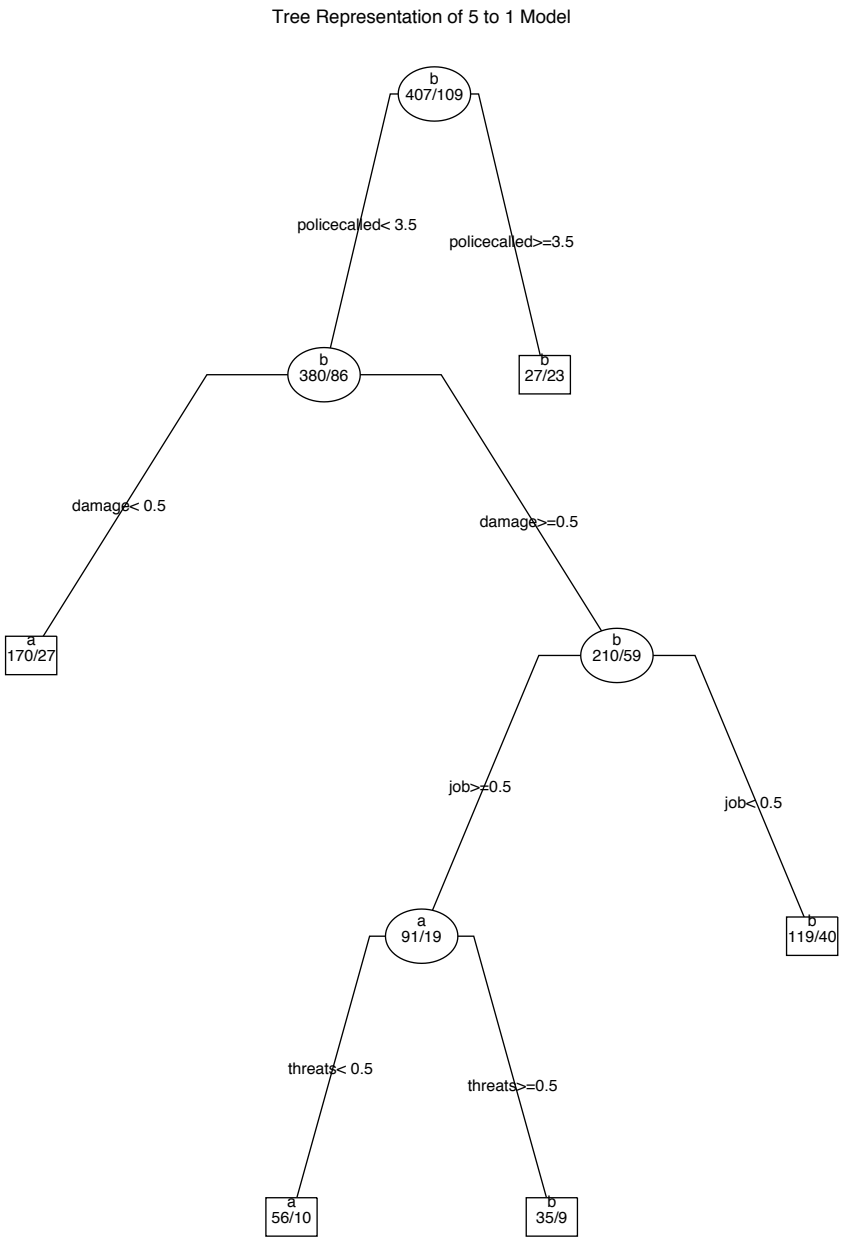


Figure 5: Classification Tree for 5 to 1 CART Model

households in the future. A prediction of a new call would be made under three scenarios as follows:

I. If the police have been called more than 3 times to a household;

II. If the police have been called 3 or fewer times, and

1. the perpetrator is reported to destroy property when angry, and
2. the perpetrator is unemployed;[5]

III. If the police have been called 3 or fewer times, and

1. the perpetrator is reported to destroy property when angry, and
2. is employed, and
3. is reported to have threatened to kill the victim or someone else in the family in the past.

The role of prior police calls is not surprising. It makes sense that past police calls is a good predictor of future police calls. Using that predictor alone with the cutoff of 4 or more past calls, forecasts of future calls could be correct nearly half the time. However, the real power of the analysis is that even when there have been *fewer* police calls in the past, there are other predictors that can be very useful. The first such predictor is whether the perpetrator is reported to do property damage around the home when angry. And the stakes are raised even higher if such individuals are unemployed. Finally, if such a person is employed, the fallback predictor is whether the perpetrator is reported to have threatened in the past to kill the victim or someone else in the family.

Note that although there were approximately 30 items in the long screener, only 4 are needed. No other items meaningfully improved the results; no other items improved accuracy if included in addition to the 4, and substituting other items for any or all of the 4 did at least a little worse.

[5]Binary predictors are coded 1 for the presence of the attribute or action and 0 otherwise. Splits at .5 or greater imply a value of 1 while splits less than .5 imply a value of 0.

5.5 Addressing Overfitting

Even though the CART results look promising, CART is vulnerable to overfitting (Breiman, 2001). In brief, all fitting procedures respond to the data provided so that if another random sample from the same population is drawn and analyzed in the same manner, the results will be different. If the statistical procedure fits a complicated function, the results often will be very different. Therefore, generalizing results from the initial sample can be risky. Moreover, the usual techniques used to evaluate the model provide no information about how serious the overfitting is. In short, overfitting can seriously compromise forecasting accuracy.

A far more realistic assessment can be obtained for a classification tree by using a procedure called "random forests" (Breiman, 2001).[6] A key is that data used to evaluate how well the model performs are *not* used to build the model. A large number of classification trees are constructed, each based on a bootstrap sample of the data. In addition, at each split a random subset of predictors is selected. For each tree constructed, data not included in the bootstrap sample are used to evaluate how well the tree performs. Finally, overall results are produced by averaging over the trees.

The use of multiple trees (often as many as 1000) makes the random forests fitting function much more complicated than the CART fitting function. However, the data not included in each bootstrap sample can provide a prudent estimate of how well the model performs, and the averaging over trees directly compensates for the overfitting itself. Therefore, the random forest results can be treated as true forecasts. The data in each bootstrap sample are used to build a classification tree, and the data held out from each bootstrap sample are forecasted. How good the forecasting is can then be directly determined. This is no different from forecasting into a new random sample from the same population.

Given the 5-to-1 cost ratio, the random forest results can be seen in Table 2.[7] Because of the sampling process, the total number of cases is slightly different, but that does not affect the interpretation. As expected, forecasting skill declines a bit. Instead of being able to forecast future calls

[6]There is a large literature on overfitting and on obtaining estimates of the amount of damage done. See, for example, Hastie et al. (2001: chapter 7).

[7]Because we are now doing true forecasts, the labels along the top of the classification table differ from those used for CART classification tables. We emphasize in the labels that CART classifies while random forests forecasts.

correctly nearly two-thirds of the time, a more accurate expectation is to be correct about 60% of the time. Likewise, the 56% figure for accurately predicting an absence of calls is more reasonably pegged at about 47%. Still this is far better than one would do ignoring the four predictors.

	No Call Forecasted	Call Forecasted	Proportion Correct
No call	190	217	0.47
Call	45	65	0.59

Table 2: Classification Table for the 5-to-1 Random Forest Model

Random forest provides an additional piece of information about overfitting. Because a large number of classification trees is constructed using random samples of the data and randomly selected sets of predictors at each step, the tree structures produced will typically vary, often substantially. When the full set of predictors includes at least some that are highly related, there can be many classification trees that forecast equally well despite having different structures.

For our data, there are a number of predictors that are highly correlated. For example, whether the perpetrator had been previously convicted is strongly related to whether that perpetrator had a previous arrest. And both are strongly related to the number of previous calls to the police. Therefore, all three variables are for our purposes measuring much the same thing, and forecasting skill does not change much no matter which predictor is used. In other words, variables that may be rather different conceptually can be almost indistinguishable empirically.

The key implication is that although the CART results reported earlier provide a good and easily implemented forecasting tool, there are a number of other trees using closely related variables that do about as well. For instance, a tree using the number of arrests rather than the number of calls to the police to define the initial split of the data, performs about as well.[8]

[8]Because of the averaging, there is no single tree structure in random forests to interpret. As an alternative, random forests provides several measures of the importance of each predictor averaged over trees. There is some controversy about the properties of these measures and precisely what they convey. A discussion of the issues is beyond the scope of this paper (see Berk, 2004), and the random forests importance measures would not affect the conclusions reached here.

For these data at least, trying to disentangle the different roles of strongly related predictors is not productive.[9]

6 Forecasting New Domestic Violence Offenses

It also is important to forecast not just any calls for the same household, but new calls that result in "probable cause" misdemeanor or felony domestic violence. These are calls for service from which deputies can establish probable cause that misdemeanor or felony domestic violence has occurred. There were 29 such events during the follow-up period, representing 5.6% of the households. We will proceed in essentially the same manner as we did for predicting new calls.

Figure 6 shows the probability distribution for a logistic regression that used all available screener items. As before, logistic regression does not perform well. There are only 4 households for which the chances of a subsequent call for misdemeanor or felony domestic violence are greater than .5.

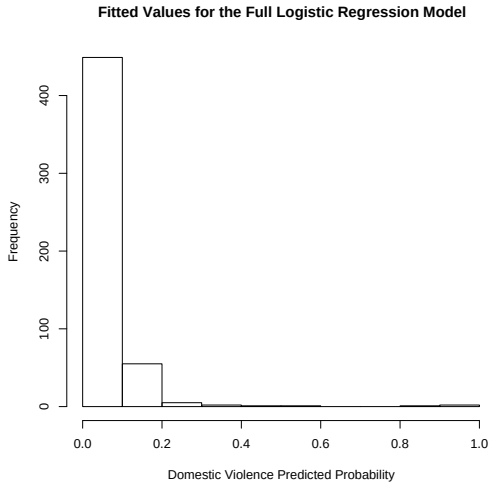


Figure 6: Probability Distribution for a New DV Offense Produced by Logistic Regression

[9]Which at least raises the question of how productive such attempts are, more generally, in research on domestic violence.

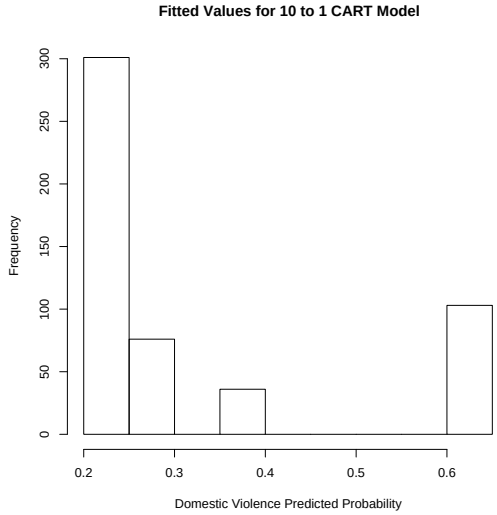


Figure 7: Probability Distribution for a New DV Offense Produced by 10-to-1 CART Model

We again turn to CART. But this time, we use a 10-to-1 cost ratio of false negatives to false positives. This, too, is consistent with information elicited from the Sheriff's department and leads to useful results. The harm of failing to forecast accurately probable cause domestic violence offenses is considerably larger than the harm of failing to forecast accurately any new calls for service. Figure 7 shows that CART generates a substantial number of households for which the chances of a subsequent domestic violence misdemeanor or felony are larger than .5.

Rare events are notoriously difficult to forecast. Table 3 shows that despite the small number of new domestic violence offenses in the 3 month follow-up period, CART does a good job of identifying them. Over 80% of the households with no new domestic violence offenses are properly identified, and 50% of the households with a new domestic violence offense are properly identified. However, because the ratio of false positives to false negatives is only a little over 6 to 1 (88/14), our results do not place quite as much weight on false negatives as the Sheriff's Department would like.[10]

[10]This results from the classification tree (see Figure 8) were stable and easily interpreted. Larger trees would have produced the desired 10-to-1 ratio but would not have

Table 4 shows that as before, CART overfits the data a bit. The forecasting skill for true positives drops from 52% to 49% while the forecasting skill for true negatives drops from 82% to 70%. But the price paid is small. Note that we now have approximately the 10 to 1 ratio of false negatives to false positives needed.

	Identified as No DV	Identified as DV	Proportion Correct
No DV	399	88	0.82
DV	14	15	0.52

Table 3: Classification Table for The 10-to-1 CART Model

	No DV Forecasted	DV Forecasted	Proportion Correct
No DV	341	146	0.70
DV	15	14	0.49

Table 4: Classification Table for The 10-to-1 Random Forest Model

Figure 8 shows the classification tree. Three useful predictors are represented: 1) whether the police are reported to have been called in the past, 2) whether the perpetrator is reported to be unemployed, and 3) whether the violence is reported to be getting worse.

A prediction of a new domestic violence offense would be made under three scenarios as follows:

I. The police have been called to the household in the past; and

II. The perpetrator is unemployed; and

III. The violence is getting worse.

In other words, there are three situations, in order of increasing likelihood, for which a future domestic violence offense should be forecast. But even the lowest level (i.e. level I) shows useful forecasting skill despite the fact that new calls involving probable cause domestic violence misdemeanors

provided a useful forecasting tool. These are the kinds of statistical tradeoffs one often faces when the event to be identified is relatively rare.

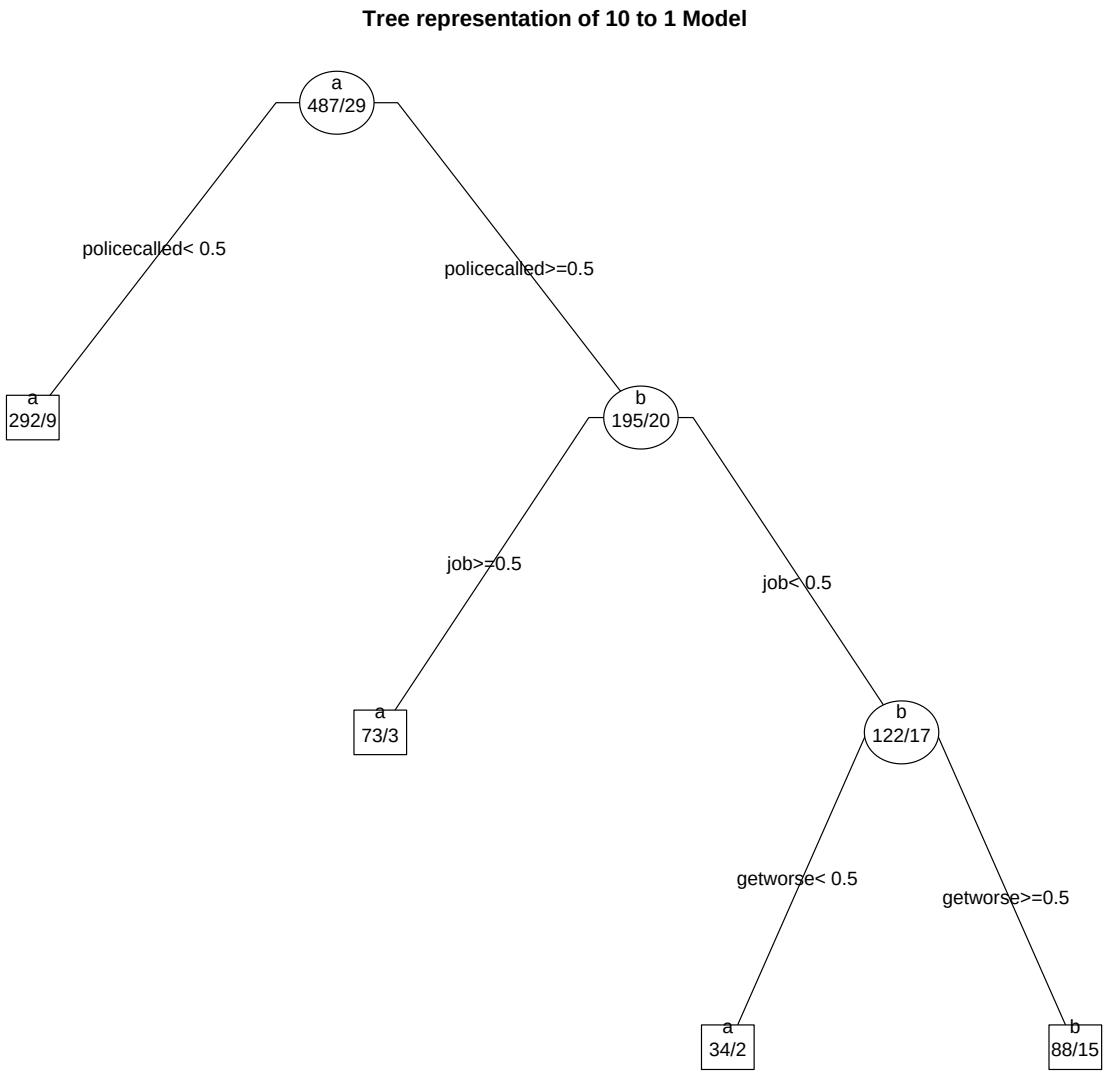


Figure 8: Classification Tree for 10 to 1 CART Model

and felonies are relatively rare. And as before, there are other trees, based on predictors that are highly correlated with the three predictors shown in Figure 8, that forecast about as well.

7 Conclusions and Policy Implications

The conclusions are straightforward. First, it is possible to forecast with useful skill future calls for law enforcement assistance. Using a cost ratio of 5 to 1 for false negatives (incorrectly forecasting no future calls) to false positives (incorrectly forecasting future calls), one can accurately forecast future calls about 60% of the time and accurately forecast the absence of domestic violence calls nearly 50% of the time.

Second, one can accomplish this with just four predictors from our larger screening instrument. These four predictors are: 1) whether the victim reports that there have been more than 3 police calls to the household before, 2) whether the perpetrator is reported to damage household property when angry, 3) whether the perpetrator is reported to be unemployed, and 4) whether the perpetrator is reported to in the past have threatened the life of the victim or someone in the victim's family. One cannot do meaningfully better by adding more predictors.

Third, using a cost ratio of 10 to 1 for false negatives to false positives, one can accurately forecast *domestic violence* calls about 50% of the time and accurately forecast the absence of *domestic violence* calls nearly 70% of the time. These are calls for which deputies can establish probably cause that a domestic violence misdemeanor or felony has occurred.

Fourth, for subsequent probable cause domestic violence calls, just three predictors are required. These three predictors are: 1) whether the police are reported to have been called in the past, 2) if the perpetrator is reported to be unemployed, and 3) if the violence is reported to be getting worse. One cannot do meaningfully better including more predictors.

However, one must keep in mind several important caveats. New calls for service do not necessarily mean that a domestic violence incident has occurred, although in most cases it probably has. Even for calls in which deputies find a probable cause domestic violence offense, the offense may later prove to be "unfounded." And there are surely a large number domestic violence incidents for which no call to the police is made. The calls for service forecasted in this analysis are significantly related to domestic violence

incidents, but are not the same thing.

In addition, although it may be tempting to infer that our best predictors are also important *causes* of domestic violence, we counsel great caution. For example, unemployment may contribute to domestic violence or, alternatively, the kinds of individuals who have trouble finding and holding jobs may tend to be the same kinds of individuals who can be violent at home. A proper causal analysis would require a different kind of study.

Finally, forecasting is necessarily data dependent. Here, it is likely that forecasting skill would change a bit if the forecasts were for either shorter or longer term outcomes. We suspect that one would forecast a bit better in the shorter term and a bit worse in the longer term. More important, only about half of the specified number of households were included in the study, and it is unclear how these were chosen. The deputies who participated in the data collection had wide discretion in when to employ the screener, and we do not know how this discretion was exercised. It is possible, therefore, that the group of households on which the analysis rests is somehow atypical. We find no evidence of this in the data, but it remains a possibility. The best solution would be to replicate the study with a sample of households known to be representative.

Given these caveats, one has to be a bit circumspect about policy implications. Nevertheless, it is it seems clear that for households prone to domestic violence it is possible to develop quick-response threat assessment instruments that can be used successfully in the field by law enforcement personnel. It seems equally clear that such instruments can take the relative costs of forecasting errors into account and then do better than a one-size-fits-all prediction. And, such instruments can be rigorously evaluated.

Whether the instruments we have developed for the Los Angeles Sheriff's Department can be effectively applied elsewhere is less apparent. Given the nature of the predictors and the strong relationships some have with each other, it is likely that the predictors selected in other sites would be a bit different from those that surfaced in Los Angeles. More important, it is entirely possible that with a different mix of households, or a somewhat different cost structure, those predictors *should* be different. We suspect, consistent with the current research literature, that the same broad kinds of variables would be relevant, but their weight and their specific measures could well vary by site. It short, it would probably be good policy for local law enforcement jurisdictions to develop their own quick-response threat assessment instruments. The main message of our study is that such an enterprise is desirable

and feasible.

References

Berk, R.A., Berk, S.F., Newton, P.J. and D.L. Loseke. (1984) "Cops on Call: Summoning the Police to Domestic Violence Incidents," *Law and Society Review* 18(3):479-498, 1984.

Berk, R.A. (2004) "Data Mining with a Regression Framework," UCLA Department of Statistics Preprint Series # 371.

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.

Campbell, J.C. et al. (2003) "Risk Factors for Femicide in Abusive Relationships," *American Journal of Public Health* 93(7), 1089-1097.

Cattaneo, L.B. and L.A. Goodman (2003) "Victim-Reported Risk Factors for Continued Abusive Behavior: Assessing the Dangerousness of Arrested Batterers," *Journal of Community Psychology* 31(4): 349-369.

California Department of Justice (2003) *Crime and Delinquency in California*, State of California, Office of the Attorney General, Bureau of Criminal Information and Analysis (www.ag.ca.gov/cjsc/index.htm)

Dutton, D. G., and R.R. Kropp (2000) "A Review of Domestic Violence Risk Instruments," *Trauma, Violence, and Abuse* 1(2): 171-181.

Efron B. and R.J. Tibshirani (1993) *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Fein, R.A., Vossekuil, B., and G.W. Holden (1995) "Threat Assessment: An Approach to Prevent Targeted Violence," National Institute of Justice, Office of Justice Programs, *Research in Action*.

Hastie, R., Tibshirani, R. and J. Friedman. (2001) *The Elements of Statistical Learning*. New York: Springer-Verlag.

Houry, D., Parramore, C. Fayard, G., Thorn, J., Heron, S. and A. Kellerman. (2004) "Characeristics of Household Addresses that Repeatedly Contact 911 to Report Intimate Partner Violence," *Academic Emergency Medicine* 11: 662-667.

Roehl J., and K. Guertin (2000) "Intimate Partner Violence: The Current Use of Risk Assessments in Sentencing Offenders," *The Justice System Journal* 21 (2) 171-198.

Sherman, L.W. (1992) *Policing Domestic Violence*. New York: Free Press.

Sherman, L.W. , Smith, D.A., Schmidt, J.D., and D.P. Rogan. (1992) "Crime, Punishment and stakein Conformity: Legal and Informal Control of Domestic Violence," *American Sociological Review* 57: 680-690.

Skogan, W and K. Frydl, eds. (2003) *Fairness and Effectiveness in Policing: The Evidence*, Washington D.C., The National Academies Press.

Straus, M.A., and R.J. Gelles (1990) *Physical Violence in American Families: Risk Factors and Adaptations to Family Violence in 8,145 Families* New Brunswick, NJ.: Transation Publishers.

Appendix A— The Long Screening Instrument

1. Is this the first time he/she has tried to hurt you? (*Circle one*)
   (a) No [1]
   (b) Yes [2] — **Skip to #12**
2. When was the last time? (*Circle the one for the most recent event*)
   (a) Earlier today [1]
   (b) Within the past week [2]
   (c) Within the past month [3]
   (d) Within the past 6 months [4]
   (e) Within the past year [5]
   (f) Longer than a year ago [6]
3. How many times before has he/she tried to hurt you?
   (a) . . . . . Times
4. How many times before have the police been called?
   (a) . . . . . Times — **If 0 skip to #7**
5. Was he/she ever arrested for domestic violence as a result? (*Circle one*)
   (a) No [1]
   (b) Yes [2]
   (c) Don't know [3]
6. Was he/she ever convicted of domestic violence as a result? (*Circle one*)
   (a) No [1]
   (b) Yes [2]
   (c) Don't know [3]
7. Is the violence getting worse as time goes on? (*Circle one*)

(a) No [1]

(b) Yes [2]

(c) Don't know [3]

8. How long ago did the violence start? (*Circle the most recent time that is appropriate*)

   (a) Within the past week [1]

   (b) Within the past month [2]

   (c) Within the past 6 months [3]

   (d) Within the past year [4]

   (e) Longer than a year ago [5]

9. Has he/she ever hurt you so that you needed to see a medical doctor? (*Circle one*)

   (a) No [1] — **Skip to #12**

   (b) Yes [2]

10. How many times?

    (a) . . . . . Times

11. Were you ever treated for those injuries in a hospital emergency room? (*Circle one*)

    (a) No [1]

    (b) Yes [2]

12. Does he/she have a problem with jealousy? (*Circle one*)

    (a) No [1]

    (b) Yes [2]

13. Does he/she keep track of whom you talk to on the phone? (*Circle one*)

    (a) No [1]

(b) Yes [2]

14. Does he/she try to determine which of your friends you can see? (*Circle one*)

    (a) No [1]

    (b) Yes [2]

15. Does he/she try to put you down in front of your friends or family? (*Circle one*)

    (a) No [1]

    (b) Yes [2]

16. Does he/she have a drinking problem or a problem with drugs? (*Circle one*)

    (a) No [1]

    (b) Yes [2]

17. When he/she is angry with you, does he/she ever try to destroy things around the house? (*Circle one*)

    (a) No [1]

    (b) Yes [2]

18. Has he ever threatened to kill you or someone in your family?

    (a) No [1]

    (b) Yes [2]

19. Are there any children in the home? (*Circle one*)

    (a) No [1] — **Skip to #23**

    (b) Yes [2]

20. How many?

    (a) . . . . . Children

21. What are their ages, starting with the youngest? (*List the ages in chronological order*)

    (a) (. . .) (. . .) (. . .) (. . .) (. . .) (. . .) (. . .) (. . .) (. . .) (. . .)

22. Has he/she ever intentionally hurt him/her/any of them just because he/she was angry? (*Circle one*)

    (a) No [1]
    (b) Yes [2]

23. Does he/she have a handgun he/she can get to? (*Circle one*)

    (a) No [1] — **Skip to #26**
    (b) Yes [2]

24. Did he/she purchase it himself/herself? (*Circle one*)

    (a) No [1]
    (b) Yes [2]

25. When he/she is angry, has he/she ever threatened you with it? (*Circle one*)

    (a) No [1]
    (b) Yes [2]

26. Does he/she have a rifle he/she can get to? (*Circle one*)

    (a) No [1] — **Skip to #29**
    (b) Yes [2]

27. Did he/she purchase it himself/herself? (*Circle one*)

    (a) No [1]
    (b) Yes [2]

28. When he/she is angry, has he/she ever threatened you with it? (*Circle one*)

    (a) No [1]

(b) Yes [2]

29. Has he/she ever threatened you with some other weapon like a knife? (*Circle one*)

    (a) No [1]

    (b) Yes [2]

30. Is there a restraining order against him/her right now? (*Circle one*)

    (a) No [1]

    (b) Yes [2]

    (c) Don't know [3]

31. Have you ever left him/her (*Circle one*)

    (a) No [1] — **Skip to # 33**

    (b) Yes [2]

32. How many times?

    (a) . . . . . Times

33. Does he/she have a regular job? (*Circle one*)

    (a) No [1]

    (b) Yes [2]