J Exp Criminol
DOI 10.1007/s11292-013-9193-0

Increased homicide victimization of suspects arrested for domestic assault: A 23-year follow-up of the Milwaukee Domestic Violence Experiment (MilDVE)

Lawrence W. Sherman · Heather M. Harris

Abstract
Objective To test for any long-term effects on the death rates of domestic assault suspects due to arresting them versus warning them at the scene.
Methods The Milwaukee Domestic Violence Experiment (MilDVE) employed a randomized experimental design with over 98 % treatment as assigned. In 1987–88, 1,200 cases with 1,128 suspects were randomly assigned to arrest or a warning in a 2:1 ratio. Arrested suspects were generally handcuffed and taken to a police station for about 3 to 12 h. Warned suspects were left at liberty at the scene after police read aloud a scripted statement. Death records were obtained in 2012–13 from the Wisconsin Office of Vital Statistics and the Social Security Death Index, with the support of the Milwaukee Police Department.
Results In the first presenting case in which the 1,128 were identified as suspects, they were randomly assigned to arrest in 756 cases and to a warning in 372. No clear difference in death rates from all causes combined (d=0.04) was ever evident between the groups, or for five of the six specific categories of cause of death. However, a clear difference in homicide victimizations of the suspects emerged between those arrested and those warned. At 23 years after enrolment, suspects assigned to arrest were almost three times more likely to have died of homicide (at 2.25 % of suspects) than suspects assigned to a warning (at 0.81 %), a small to moderate effect size (d=0.39) with marginal significance (two-tailed p=0.096; relative risk ratio=2.79:1; 90 % CI = 1.0007 to 7.7696). Cox regressions controlling for suspects' stakes in conformity (employment and marriage) show that homicide victimization for arrested suspects is three times that of warned suspects (p=0.07), although no interactions are yet significant. Logistic regression with more covariates increases arrest effects on homicide to 3.2 times more than warnings (p=0.06).

This article is dedicated to the memory of Anthony Bacich and Dennis Rogan, who worked hard to make this experiment a clear and powerful test.

L. W. Sherman (✉) · H. M. Harris
University of Maryland, College Park, MD, USA
e-mail: shermanl@umd.edu

H. M. Harris
e-mail: hmh@umd.edu

Published online: 05 October 2013

Conclusions Suspects randomly assigned to arrest died from homicide at a consistently higher rate than controls over a two-decade period, but the difference was not statistically discernible until the 22nd year after assignment. Long-term follow-up of randomized experiments is essential for detecting mortality differences that substantially affect cost–benefit analyses of criminal justice practices.

Keywords Arrest · Domestic violence · Policing · Randomized experiment · Mortality-defiance theory

Introduction

In the first multi-decade follow-up study of a randomized field experiment that attempted to prevent crime, McCord (1978; 1981) discovered important effects beyond the originally planned outcome criteria. For the intended criteria, she found evidence of more serious criminality over 30 years among young boys who had received a mentoring and services program in comparison to untreated controls. However, on criteria beyond the original design of the experiment, she found even clearer evidence of harm. Most notably, she reported that among those who died in the 30-year follow-up period, those who received the program died at a mean age of 32, compared to a mean of age of 38 years at death among control group members who had died—a 21 % reduction in life span (McCord 1978: 286).

A modern critique of her approach could ask whether it was a "fishing expedition" in which she looked at so many outcomes from the same data set that the death result could have been a "false discovery" that occurred by chance, especially since the prevalence of death was the same in each treatment group by 30 years after random assignment. McCord (1978: 288) examined that possibility by showing that the seven comparisons out of 57 that favored the controls had only a 2 % risk of occurring by chance at p=0.05. Yet quite apart from her overall measures, and notwithstanding the debate about best statistics to check for "false discovery" in multiple significance tests (e.g., Benjamini and Hochberg 1995), there are good grounds to examine death even without other measures.

Death is fundamentally different from any other outcome. Nothing measures a life-course as fundamentally as its length. While McCord's measures of alcoholism, mental health and job satisfaction, for example, were vulnerable to measurement error and incomplete responses, her measurement of death was comprehensive and certain as to its meaning. In her study, and arguably all longitudinal studies of experimental effects, death should arguably be an essential measure for fully understanding the consequences of treatments—even if it only to control for time at risk of other outcomes. Whether or not it was written up for an initial experimental protocol, death is an essential dimension of cost–benefit analysis of life-course treatment impact.

Death is the focus of this article, written in the context of a series of analyses of our long-term follow-up of an experiment that was not originally focused on death as an outcome. The experiment was designed to measure what was originally reported: the effect of arrest on repeat domestic violence in the short run (Sherman et al. 1991). This article addresses a different question: *what effect did arrest have on the death rates of the suspects*, *and how long did it take before there was adequate power to*

discern any differences arrest may have caused in suspect mortality? The second half of this question has methodological implications for all policy experiments in terms of cost–benefit analyses for the policies tested. The first half of the question has substantive implications for the accumulating evidence on the costs and benefits of arrest for misdemeanor domestic assaults. We return to both these themes in the conclusions of the article.

In the spirit of McCord's pioneering work, in 2011–13 we collected data on both the criminality and death rates—but nothing else—of both offenders and victims in the Milwaukee Domestic Violence Arrest Experiment (MilDVE). Our research comprises the first 23-year follow-up of any of the five completed replication-extensions (Maxwell et al. 2002) of the Sherman and Berk (1984) Minneapolis Domestic Violence Experiment (often cited as the "MDVE, but which we shall call the "MinDVE" to distinguish the Minneapolis experiment from its methodologically improved replication in Milwaukee in 1987–88). The present report is one of our series of papers that uses different analytic frameworks to examine distinctive questions about long-term effects of arrest for misdemeanor domestic violence.

In one paper in this series (Sherman and Harris 2013a), we present evidence that arrest had no long-term, discernible effect on repeat domestic violence. We also present evidence from a pre-planned moderator analysis (Sherman 1984) that among suspects who had been unemployed when randomly assigned, arrest significantly increased recidivism over most of, and at the end of, the 23-year period. We also found no deterrent effect of arrest among employed suspects (after the first few years), contrary to the initial findings (Sherman and Smith 1992). These two-decade findings offer a far longer view of recidivism outcomes than any of the prior experimental evidence, which has been limited to 2-year follow-ups of both official data and victim interviews (Berk et al. 1992; Pate and Hamilton 1992; Sherman 1992; Sherman et al. 1992; Maxwell et al. 2002). Notably, the long-term findings contradict the short-term evidence from a pooled analysis of the six experiments that showed arrest had a significant main deterrent effect on domestic violence. The new findings also contradict one aspect of defiance theory (Sherman 1993), which predicted incorrectly that sanctions would deter crime among those who have greater stakes in conformity.

In another paper in this series (Sherman and Harris 2013b), we report a modestly but significantly higher death rate from all causes among all victims whose offenders were arrested (compared to controls), with heart disease as the leading cause of death, while almost no victims died of homicide in either treatment group. The effect size of suspect arrest increasing 23-year victim death rates is more than doubled among victims who were themselves employed. These negative effects of suspect arrest on victim mortality are robust, surviving tests employing a variety of controls and survival models. The victim findings are consistent with a range of medical and biological evidence about the impact of a single, brief, but stressful event on morbidity and mortality.

In the present and third paper of the series, we examine the death rates of the suspects, which show very different patterns from the death rates of the victims. Unlike the MilDVE victims, the suspects in the arrest group experienced no increase in overall death rates. Unlike the victims, the MilDVE suspects experienced no higher death rates from heart attacks in the arrest group than in controls. And unlike the victims, the MilDVE suspects *did* suffer a higher death rate from homicide in one treatment condition than in the other.

We decided to present the evidence in separate papers, in part, due to the complexity of the data. More importantly, however, our results suggest that the etiology of mortality differs for domestic violence offenders and domestic violence victims. Despite efforts to unify theories of victimization and offending with respect to criminal offending (e.g., Lauritsen et al. 1991), our findings suggest that different theoretical frameworks are required to interpret suspect and victim mortality. Rather than forcing the data analysis into any unified deductive framework, we choose to explore the findings inductively, question by question.

The present paper examines its evidence with the same inductive approach adopted by McCord (1978, 1981). Like the Cambridge-Somerville experiment, the MilDVE was not designed to test the effects of arrest on suspect death rates. There is no victimization theory, to our knowledge, about the effects of arrest on the subsequent risk of being murdered, just as there was no deductive theory that could have predicted that a mentoring program for young boys would affect their longevity. Like McCord (1978), we examine the data at hand and consider its implications for theory and policy.

Policy context

The substantive policy context of this discussion is about the advent of mandatory arrest policies for misdemeanor domestic assault (not witnessed by police) in the mid-1980s. One key question in the debate about the value of such policies is the role that short-term experimental findings played in causing them to be adopted. Several scholars have attributed the sudden adoption of mandatory arrest policies for misdemeanor domestic violence to the Minneapolis Domestic Violence Experiment (Sherman and Berk 1984), one of the top 50 most-cited articles in the first 100 years of the American Sociological Review (Jacobs 2007). Iyengar (2009: 87), for example, notes that

> "…arrest policies might not have been so widely adopted had it not been for … the Minnesota Domestic Violence Experiment (MDVE). The results of this experiment were used by the US Department of Justice, academics, legislators, and criminal justice spokespersons to justify and support mandatory arrest policies. For example, the Violence Against Women Act (1994) used the results from this experiment to justify grants and funds to support pro-arrest policies in various states."

Dugan et al. (2003: 172) report more systematic evidence of

> "widespread adoption of pro-arrest policies after …Minneapolis research indicating that arresting the batterer reduces the chances of continued partner violence. Data collected by the authors show that prior to the mid-1980s few jurisdictions had proactive arrest policies, yet beginning in 1984 the trend of aggressive arrest policy rose dramatically."

Many scholars have expressed great concern about the influence of a single experiment having such a broad policy impact without replication (e.g., Lempert 1989; Sherman and Cohn 1989; Dugan et al. 2003; Sampson et al. forthcoming). However, what is rarely mentioned is potentially a greater concern than the lack of replication: the

brief follow-up period for measuring the effect. The Minneapolis experiment used only a 6-month window of subsequent behavior of the suspects to reach its conclusions. While fairly common practice for evaluation research at that time, a 6-month window offers a very limited view of the costs and benefits of a policy for criminal sanctions. Even the five replications of Minneapolis were limited, largely by budgetary constraints, to 2 years or less as the observation period for the outcome measures.

There are many obstacles to using longer-than-2-year follow-up periods for evaluating policies that can be tested for impact on the individual life course. Funding bodies such as the National Institute of Justice rarely make grants for more than 2 or 3 years. Few grants are ever awarded for long-term follow-ups. Few policies or Institutional Review Boards would allow preservation of individual identifiers for the decades necessary to allow such follow-up. Unless the original principal investigator keeps the records secured in their homes (as the late Joan McCord did) or in the Harvard Law Library (as Eleanor and Sheldon Glueck did), there may be no way to conduct a long-term follow-up of mortality or anything else. Thus, long-term follow-up is rare. However, it is arguably essential to achieve a full understanding of the effects of a policy choice.

With the policy debate over mandatory arrest still continuing in both the UK and US, we decided that a long-term follow-up would contribute additional facts for discussion. In this paper, we do not assess the details of the debate. Rather, we provide new facts derived from the identifiable data that was securely preserved by Sherman since the end of the MilDVE. With the support of Chief Edward Flynn of the Milwaukee Police Department, we gathered data on arrest histories of both suspects and victims and police reports pertaining to homicide victims. The Wisconsin Office of Vital Statistics supplied information on mortality by cause of death.

Theoretical context: How long is long "enough" to determine costs and benefits?

A key theoretical issue for evidence-based policy may be the assumption that most effects of criminal justice are transitory, and therefore only short-term evaluations are needed. Even in medicine, 5-year follow-up periods are common, but that may be because most patients are dead 5 years after many treatments are tested. Yet the true cost–benefit picture of a treatment may not be revealed until three or four decades after random assignment, as illustrated by such examples in experimental criminology as the Perry Pre-school Experiment (Nores et al. 2005) and Joan McCord's (1978) 30-year follow-up of the Cambridge-Somerville delinquency prevention program. A comprehensive cost–benefit analysis of mandatory arrest policies is beyond the scope of this paper, but any such analysis must incorporate any long-term effects such as those we report.

As of mid-2013, not one of the six NIJ-funded experiments in arrest for misdemeanor domestic assault had been the subject of a published follow-up on any outcome indicators for more than 2 years. The benefit of doing such follow-ups may be twofold. One is that initially detected effects may be found to have persisted or disappeared. The other is that effects on rare events that were not initially discernible may become so over time, as the statistical power of the test increases with more time at risk for all of the sample members to accumulate rare events. In this paper, we explore the latter possibility as we focus on death as an additional outcome measure.

Two prior studies have examined the macro-level effect on populations of mandatory arrest policies for misdemeanor domestic assault on rates of death from domestic homicides (Dugan et al. 2003; Iyengar 2009). The studies used different designs and reached opposite conclusions on whether mandatory arrest reduced or increased homicides of abused women. Neither used experimental designs, nor examined death rates in the life course of individuals. That does not make their findings unimportant, since individual-level experiments cannot answer questions about macro-level effects, but it does mean that the individual-level effects remain unknown, both on homicide and other causes of death. In a separate paper, as noted above, we report that there is no individual-level effect of suspect arrest on long-term homicide rates of the domestic assault victims in our sample, but arrest did cause an increase in overall victim mortality. In this paper, we examine all causes of death among suspects.

Why should arrest affect the mortality of domestic misdemeanor assault suspects?

In asking whether it is plausible that arrest for domestic misdemeanor assault could affect death rates of suspects arrested, there are two relevant theoretical frameworks, biological and criminological. A *biological* framework would predict death by cardiovascular and stress-related causes of death, but not a difference in homicide. A *criminological* framework would predict an increase in homicide, but not in other causes of death.

The biological framework predicts that arrest is a highly stressful, traumatic experience that, left untreated, could cause elevated risks of coronary disease and general immune system failures (Kubzansky et al. 2007), and possibly depression and suicide (Oquendo et al. 2003). The negative consequences for arrest across observed *on* many other dimensions of the life course, such as labor force participation (Bushway 1998), would be consistent with the medical evidence on the effects of such traumas, even if they were minor in character. Kubzansky and colleagues (2007), in fact, report increased mortality not among those with clinically diagnosed post-traumatic stress disorder (PTSD), but among those with merely elevated symptoms of such stress that are measured at levels below those used to set the threshold of PTSD.

A criminological framework for arrest can adapt a theory of victimization to predict higher mortality from murder. That theory is the "*victim-precipitation*" hypothesis. As Wolfgang (1958) and others have observed, many homicides are provoked by aggression of the victim against the victim's killer. This aggression may be a crime by the about-to-be-murdered person against someone who is initially their victim, but about to become their murderer. Thus, their murder becomes a side effect of their increased aggression, or reduced self-control, which may be causally linked to the police decision to arrest them.

Alternatively, under Agnew's (2006) General Strain Theory, homicide could be higher for anyone who is arrested. Arrest may be a prime example of the kind of "negative life event" that contributes to cumulative strain in a non-linear fashion. Even if police conduct in the arrest is not *objectively* unfair, disrespectful or abusive, it could be perceived as *unjust* by some arrestees, as Paternoster et al. (1997) found with the Milwaukee arrestee interview data. The idea of accumulated frustration

building up over long time periods (as well as in the short term) has an extensive theoretical tradition (e.g., Bernard 1990). On the premise that each arrest for domestic violence could increase frustration and aggression, anyone who is arrested may become more likely to provoke murder.

Whether the increase in provocation is linear or geometric is a key theoretical issue, beyond the scope of this paper. For this explanation to fit the facts, it would require the assumption that a single arrest can create a "tipping point" for a sharp increase in risk of the individual's behavior becoming so provocative that they will be much more likely to "cross over a line" of acceptable conduct, even among friends (Anderson 1978).

Chaos theory provides a mathematical paradigm for the idea of a non-linear accumulation of frustration—what may be most powerful about a GST prediction of more homicide from arrest. Given the short duration of an arrest experience in a lifetime of many negative life events, the impact of an arrest on life outcomes seems unlikely to carry much weight in a linear model of causation. Yet from a "chaos theory" (Gleick 1987) standpoint, the small initial difference between arrest and a police warning could plausibly trigger a slight divergence in life course pathways. This divergence could become magnified over time, with increased anger and reduced self control (Agnew 2006: 63) in the arrested group becoming larger over time, relative to the warned group. By the end of the 23-year follow-up, the small initial differences could lead to much larger differences in behavior that could affect mortality from many causes. Thus, the frustration generated by a single arrest could make the behavior of those in the extreme tail of the distribution far more provocative in the arrest group than in the warned group, leading to more frequent victim-precipitated homicide (Wolfgang 1958).

Finally, defiance theory (Sherman 1993) suggests that offenders who have fewer stakes in conformity, such as employment and marriage, may be more likely to commit crimes if they are punished (in their minds) illegitimately than if they are not sanctioned at all by the criminal justice system. We have already found support for that prediction in this sample in a separate paper on the long-term increase in domestic violence among unemployed suspects—although no support for the obverse prediction that employed people would be deterred by arrest. To the extent that defiance theory predicts an increase in anger from arrest, a similar causal pathway could make suspects more provocative in their behavior to those who might murder them. By extension, then, we might predict that a suspect's stakes in conformity could moderate the effect of arrest, especially by increasing homicide victimization for those who are unemployed and unmarried at the time of random assignment.

Research design, data, and methods

Study site

The Milwaukee Domestic Violence Experiment has been reported in detail in prior publications (Sherman 1992; Sherman and Smith 1992; Sherman et al. 1991, 1992). It was conducted in four districts of the Milwaukee Police Department in 1987–88. With a 1990 population of approximately 625,000 spread over 97 square miles, Milwaukee at the time of the experiment was rated as the most hyper-segregated city in the US (Massey and Denton 1993). Thirty percent of the city was African

American, but only 6 % was of Hispanic ethnicity. Although they comprised the majority of the city's population, almost no white people could be seen on the street in large areas of the city where unemployed heads of household exceeded 50 % and many single-family detached houses had been converted into two or three flats.

At the time the experiment began, there was great local interest in arrest for domestic violence, and the state legislature was considering the mandatory arrest law it later adopted. Sherman selected the city for the experiment partly based on field observation of a great volume of cases coming to police attention each night, with relatively high concentrations in small areas. The logistics of conducting an experiment under those conditions were especially favorable for the use of a small team of highly selected, trained, and managed officers, which is exactly what the late Police Chief Robert Ziarnik provided through his leadership team. He and his strongly committed staff inspector assigned a charismatic police lieutenant to manage the team on a full-time basis, until the goal of 1,200 eligible cases had been randomly assigned to the arrest and warning treatments. What they delivered was no different from what a less selective group of officers would do; they just did it more consistently with random assignment than the officers in any of the other five arrest experiments, delivering 98.5 % of cases with treatment as assigned, compared to a mean of 87 % in the others (Sherman 1992: 141).

Description of the intervention

From April 7, 1987 to August 8, 1988, MPD assigned 36 police officers to special duty in taking all domestic disturbance calls during evening hours in their districts, subject to their availability when the calls were received at headquarters. Eligible cases were those in which two uniformed police officers in Milwaukee (Wisconsin) were given "probable cause" to believe sufficient evidence that two people were (1) both present, (2) in a domestic relationship, and (3) one had just committed a misdemeanor assault against the other. Cases of serious injury were ineligible, as were attempts to inflict serious injury. Both suspects and victims were required to be older than age 18. Suspects already named in an arrest warrant for a previous alleged offence were excluded. Thirty-eight percent of the victims in the eligible cases had no visible signs of injury, in contrast to 68 % of suspects. Injuries were generally minor, with only 6 % of victims being treated at a hospital.

A minimum of two police officers arrived together in response to all domestic disturbance calls, with more pairs joining them in some cases. Police immediately separated the victim and suspect, then usually interviewed both as well as any witnesses present. The officers always performed a risk assessment for high levels of anger, references to weapons, threats or other facts that could make the case unsuitable for a warning. When they determined that there was probable cause to believe that an apparently adult eligible suspect had committed a misdemeanor assault against an adult domestic partner, the officers called the warrant desk to ensure there were no outstanding arrest warrants issued against the suspect. If the check for warrants was negative, one police officer then called the research office to supply names, addresses and dates of birth of both victim and suspect.

When cases were assigned to arrest, there was also random assignment to the speed with which arrestees were to be released from police headquarters. All arrest

cases were pre-assigned by the quasi-random sequence to be either (1) held until the next court day for arraignment (for a mean of 12 h but a maximum× for 40, or (2) released on recognizance without money bail (for a mean of 3 h). Initial analysis found little difference in outcomes between cases with "long" and "short" arrest. We therefore combined the cases in the two categories for purposes of long-term follow-up of the effect of any arrest at the scene. Police procedures that victims and bystanders witnessed at the scene in both versions of the arrest treatment were identical, including handcuffing the suspect (Sherman 1992: 330).

When cases were randomly assigned to disposition by warning, police were instructed to read aloud from a piece of paper the following warning: "1. We're not going to arrest anyone here tonight. 2. If we have to return, someone will go to jail. 3. Here is a list of people that can help you, both of you, with your problems. 4. The D.A.'s Office is on that list and you can contact them if you want to press charges." The "list" was a handout officers were required to give to all domestic violence victims, including those whose partners had been arrested. A similar script was provided for arrest cases, but observations suggested it was used less often; the script in the warning cases was almost always delivered (Sherman 1992: 402). In less than three cases, the warning was followed by the suspect behaving in a way that required police to arrest the suspect. These protocol violations were limited to suspects who became verbally belligerent or assaultive toward police or the victim.

Experimental design and implementation

Suspects were assigned by police authority to be arrested or warned (ratio 2:1) without consent of either victims or suspects, consistent with ethical standards and law for experimentation in criminal sanctions. The police at the scene called a research office in police headquarters to supply all identifying details for both the victim and suspect.

After a team of mostly recent graduates in criminology recorded the identifying details onto a computer record, they opened one of the 1,200 sequentially numbered, opaque, sealed (with sealing wax) envelopes containing the randomly assigned dispositions. The sequence of cases assigned to arrest and warnings was generated by Dr. Kinley Larntz, a Professor of Statistics at the University of Minnesota, who was independent of the research team and who periodically audited the assignment of cases as randomized. After opening the envelope, the research staff member communicated the assigned treatment in each case back to police at the scene, who remained on the phone after providing identifying details. Because randomization involved allocation to different legal statuses, it was impossible to mask research assistants, treating police, or victims from the nature of the treatment. Daily oversight by the local research manager provided unannounced auditing of the sequence of the envelopes opened to prevent opening several envelopes to find a preferred disposition.

Ethical approval for the random assignment of arrest was granted, upon recommendation of the late Milwaukee Police Chief, Robert J. Ziarnik, by a unanimous vote at a public legislative session of the Milwaukee City Council in 1986, and by an Institutional Review Board of the Crime Control Institute of Washington DC the same year. Approval for the 23-year follow-up study was obtained, with support of Milwaukee Police Chief Edward Flynn, from the Institutional Review Board of the University of Maryland at College Park in 2012, title 334834–2.

Outcome measures

The outcomes in the present paper are limited to deaths, with cause of death provided by the Wisconsin Office of Vital Statistics (OVS). Additional deaths were identified beyond OVS records by multiple searches of the Social Security Death Index (SSDI). Descriptions of homicides were provided in short narrative summaries by the Milwaukee Police Department.

Analytic techniques

Our trial randomly assigned arrest or warnings to cases, rather than to people (Fig. 1). The present research question, however, can only focus on people. The design is slightly complicated by the re-enrolment of about 5 % of the sample members in more than one of the randomly assigned cases, with 26 individuals classified as both suspects in one case and as victims in another. Our decision to allow reenrollment was based on a primary concern with short-term repeat offending, as well as on police preference to complete the project as soon as possible. For purposes of assessing long-term differences in mortality, however, multiple treatments complicate the design.

Our procedure is to apply the random assignment of intention-to-treat in the *first* case in which each individual appeared as a suspect (Fig. 2). This is arguably little different from including suspects who may have been arrested on other occasions after the experiment concluded in 1988—a point that bears heavily on the discussion of the SUTVA (stable unit treatment value assumption) issues discussed below. There are many sources of noise that complicate the interpretation of experimental effects over many years, including both prior and repeat arrests. With random assignment, however, there is a strong signal of manipulated causation that may potentially reveal independent effects even while holding constant noise from all other causes.

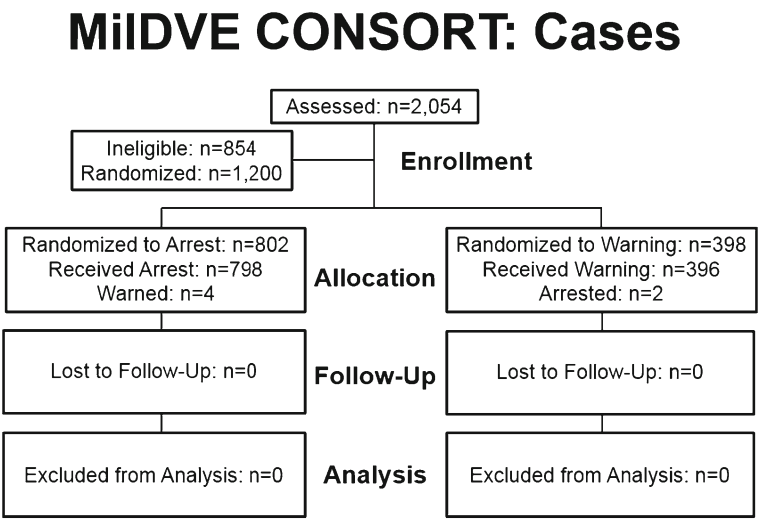

Fig. 1 MilDVE consort diagram: Cases as treated

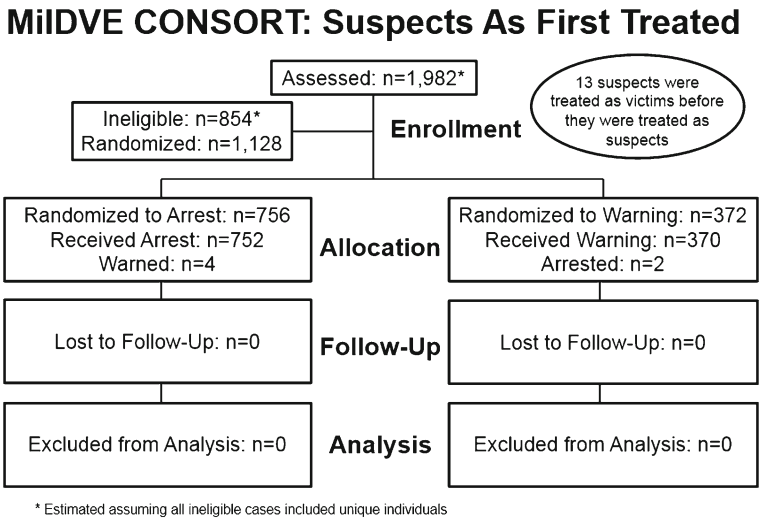

Fig. 2 MilDVE consort diagram: Suspects as first treated

As a sensitivity analysis, we also examined the results using an "ever-arrested" criterion, rather than a "first-time-as-suspect" random assignment to arrest. Using this alternate criterion changed the treatment condition for just 13 suspects. The substantive results presented below were unchanged.

Results

Of 2,054 cases assessed and 1,200 cases randomized, 1,128 people were identified as suspects in one or more cases; 26 of these were identified as both a victim in one case and a suspect in another. As Table 1 shows, the suspects were 90 % male, 75 % African American, 44 % employed, with a mean age of 32; 64 % had previously been

Table 1 Suspect characteristics at baseline

	Missing		Arrested		Warned		*t* test (α=0.10)	*p* value
	n	%	*n*	%	*n*	%		
Treatment	0	0.00	756	67.02	372	32.98	na	
Male	0	0.00	683	90.34	337	90.59	−0.1327	0.89
Black	0	0.00	572	75.66	279	75.00	0.2424	0.81
Had a job	15	1.33	365	48.28	129	34.68	−4.3275	0.00
Partner had a job	3	0.27	224	29.63	106	28.49	−0.3122	0.75
Had a prior arrest	0	0.00	503	66.53	218	58.60	−2.6136	0.01
Graduated high school	75	6.65	281	37.17	96	25.81	−5.0563	0.00
Had ever been married	2	0.18	240	31.75	106	28.49	−1.0992	0.27
	Mean	SD	Mean	SD	Mean	SD		
Age	na	na	31.57	8.65	31.63	9.05	0.1083	0.91

arrested. In the first presenting case in which the 1,128 were identified as suspects, they were assigned to arrest in 756 cases and to a warning in 372. Suspect characteristics were not entirely equivalent at baseline, with three out of eight examined showing significant differences. Biggest differences are moderately higher rates of employment and high school graduation among those arrested. Subsequent analyses therefore employ controls for imbalances.

Table 2 shows that 23 years after random assignment, suspects assigned to arrest were almost three times more likely to have died of homicide (2.25 %) than suspects assigned to a warning (0.81 %), a moderate effect size (d=0.39) with marginal significance (two-sided p=.096; relative risk ratio=2.79:1; 90 % CI = 1.0007 to 7.7696). No clear difference in overall death rates (d=0.04) was evident, nor was any cause of death other than homicide clearly different in a main effect. Thus, unlike the domestic violence victims, the suspects have no evidence of the biological pathway for arrest effects on death. What the evidence does suggest is that a criminological pathway leads from arrest to higher homicide rates.

The homicide rates in our suspect sample overall are extremely high, even by US standards, where the homicide rate in 1988 was 8.4 per 100,000 (N.Y. Times, August 13, 1989). The annualized homicide rate in our population over the 23 years is 77 per 100 K. That is almost eight times the US rate, and almost twice the average police-recorded homicide rate in Milwaukee from 1987–2011 of 38/100 K. The suspects' overall homicide rate was also 51 % higher than the mean rate in the four districts where the experiment was conducted of 51 per 100,000. This contrast merely indicates that the misdemeanor domestic assault offenders are at double the risk of getting murdered than the city-wide population. The statistical power from that high base rate is what makes it possible for us to detect an effect of randomly assigned treatment. It is in that context that arrest increases that already-high rate by almost three-fold.

Nonetheless, the significance level of the raw difference in homicide death rate between arrested and warned suspects is marginal. Given the imbalance in the baseline equivalence of the groups, we may detect a more clearly discernible effect if we control for theoretically relevant predictors of homicide victimization. Most relevant may be the idea of "stakes in conformity" that predicted different short-term recidivism in reactions to arrest (Sherman 1992, 1993). Table 3 shows that adding these covariates to a Cox regression hazard model of survival raises the effect to a full three times higher risk of homicide victimization, and lowers the p value to 0.07.

Figure 3 graphs the survival curves from the Cox regression model that is derived from Table 3.

Using even more theoretically relevant covariates to compensate for imbalances at baseline our logistic regressions show that the odds of being murdered are approximately 3.2 times higher for arrested suspects (p=0.06). This model adds in several indicators of General Strain Theory (Agnew 2006), including the idea that when a victim was employed there was even more frustration for the suspect, regardless of the suspect's employment status at the time of random assignment. Our regressions successively control for stakes in conformity, suspect demographics, and victim employment. Standard errors are clustered on individuals, to account for suspects who were treated more than once. Our result with respect to arrest is robust to specification (Table 4).

Table 2 Main effects at 23 years on suspect mortality by cause of death

Cause of death	No. of 756 arrested who died	No. of 372 warned who died	Percent arrested who died	Percent warned who died	Risk ratio	90 % CI	95 % CI	Cohen's *d*	Fisher's exact two-sided *p*	Deaths per 100 K arrested	Deaths per 100 K warned
All causes	103	47	13.62	12.63	1.078	0.823–1.413	0.782–1.488	−0.040	0.709	13,624	12,634
Homicide	17	3	2.25	0.81	2.788	1.001–7.77	0.822–9.455	−0.390	0.096	2,249	806
Heart disease	25	14	3.31	3.76	0.879	0.513–1.506	0.462–1.67	0.064	0.730	3,307	3,763
Cancer	16	10	2.12	2.69	0.787	0.409–1.515	0.361–1.718	0.119	0.534	2,116	2,688
Alcohol or drugs	11	8	1.46	2.15	0.677	0.317–1.443	0.274–1.668	0.198	0.461	1,455	2,151
Suicide	4	1	0.53	0.27	1.968	0.314–12.344	0.221–17.548	−0.277	1.000	529	269
All other causes	30	11	3.97	2.96	1.342	0.759–2.374	0.68–2.648	−0.136	0.499	3,968	2,957

Table 3 Cox regression model with suspect employment at random assignment and suspect ever married at baseline

Variable	Hazard ratio	Robust standard error	*p* value
Arrested	3.0417	1.8813	0.07
Black	0.5460	0.2583	0.20
Had a job	0.4316	0.2652	0.17

Table 5 shows the moderator analysis from an intention-to-treat design. The only clear moderator effect (one out of eight, or 12 % of those tested) on arrest causing higher rates of suspect homicide found if the suspect's victim was employed at that time (d=0.71; p=0.06). However, it is also worth noting that fully 15 of the 17 arrested suspects (88.2 %) who were later murdered had a prior arrest record, as compared to 63.9 % of the entire sample and 33 % of the arrest treatment group. Most important theoretically may be the observation that the only two categories with a complete absence of homicide victimization among those who were not arrested, *and* either employed *or* married. Stakes in conformity and an absence of criminal sanction combine to support at least a partial link between offending and victimization for both GST (Agnew 2006) and defiance (Sherman 1993) theories.

Table 6 addresses our methodological objective to demonstrate how long it takes for substantial and statistically discernible differences in mortality to appear between two treatment groups—if at all. Table 6 shows that even a 20-year follow-up period would have failed to detect an important difference in suspect homicide victimization. The year-by-year distribution of homicide arrests in both treatment arms shows that the direction of the difference never changed at all, and the magnitude of the difference changed very little after an early peak in years seven to eight at a relative risk ratio of 4.4:1, returning to around 2.5: 1 for most of the remaining follow-up

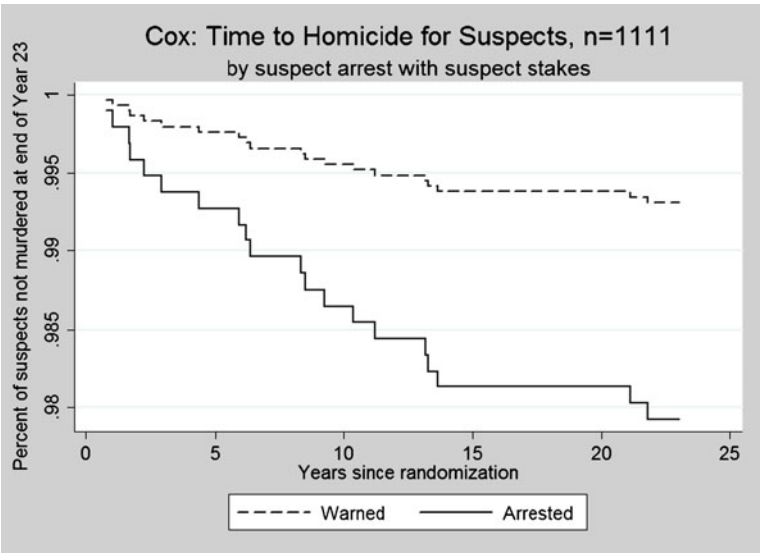

Fig. 3 Cox proportional hazards model showing homicide deaths of arrested versus warned suspects with suspect stakes in conformity controls through year 23 post-randomization

Table 4 Logistic regressions of the effect of arrest versus warning on subsequent suspect homicide, with clustered standard errors and odds ratios

	1	2	3	4	5
Arrested	3.1950	3.4367	3.2542	3.2057	3.2322
Standard error	2.0238	2.1627	2.0347	2.0098	2.0413
p value	0.07	0.05	0.06	0.06	0.06
Suspect has job		0.4972	0.5312	0.6052	0.4604
Standard error		0.2411	0.2574	0.2889	0.2270
p value		0.1500	0.19	0.293	0.116
Suspect ever married		0.3894	0.3981	0.4682	0.4114
Standard error		0.2412	0.2470	0.3083	0.2665
p value		0.1280	0.14	0.249	0.17
Suspect has prior arrest			3.1015	2.8458	2.9658
Standard error			1.9613	1.777948	1.880913
p value			0.07	0.09	0.09
Suspect race				0.2123	0.1857
Standard error				0.2182	0.1904
p value				0.13	0.10
Suspect age at assignment				0.9893	0.9874
Standard error				0.0309	0.0328
p value				0.73	0.70
Victim has job					3.1226
Standard error					1.4518
p value					0.01
Constant	0.0732	0.0114	0.0049	0.0081	0.0065
Standard error	0.0439	0.0071	0.0043	0.0081	0.0066
Observations	1,200	1,183	1,183	1,183	1,180
Suspects	1,128	1,111	1,111	1,111	1,108

period before rising to 2.8:1 at 23 years (see Fig. 4). Moreover, *the confidence intervals continued to shrink over the entire follow-up period*. All three of these facts increase our confidence in the causal relationship of the arrest treatment to the suspects' homicide victimization rates. Yet it is only at the end of the 22nd year of follow-up that the confidence intervals around the point estimates shrank enough so that the difference becomes statistically discernible at even a 10 % level.

False discovery? The repeated significance tests in Table 5 might make some readers wonder about whether the significant result after 23 years is an example of "false discovery," a chance finding that is an artifact of calculating so many significance tests. Indeed, if we calculate a Bonferroni correction (using Stata's smileplot command in multproc) for all 22 tests beyond the initial test with a standard of p=0.05, the critical value required is 0.002, or nowhere near the 10 % level we report. That kind of analysis, however, misses the central point about the problem of false discovery (Benjamini and Hochberg 1995).

Table 5 Moderator analysis of effect of arrest on suspects' homicide victimizations

Moderator	*n*	Arrested (*n*)	Warned (*n*)	Arrested killed (*n*)	Warned killed (*n*)	Risk ratio	90 % CI	95 % CI	Cohen's *d*	Fisher's exact two-sided *p*
Main effect	1,128	756	372	17	3	2.788	1.001–7.77	0.822–9.455	−0.390	0.096
Suspect never married	780	515	265	14	3	2.401	0.850–6.787	0.696–8.282	−0.353	0.198
Suspect ever married	346	240.	106	3	0	na	na	na	−0.672	0.556
Suspect unemployed	619	382	237	11	3	2.275	0.786–6.584	0.641–8.07	−0.355	0.268
Suspect employed	494	365	129	6	0	na	na	na	−0.603	0.347
Victim unemployed	795	532	263	8	3	1.318	0.436–3.987	0.353–4.928	−0.125	1.000
Victim employed	330	224	106	9	0	na	na	na	−0.712	0.062
No suspect prior arrest	407	253	154	2	1	1.217	0.164–9.063	0.111–13.314	−0.094	1.000
Suspect has prior arrest	721	503	218	15	2	3.250	0.949–11.132	0.75–14.092	−0.413	0.112

Table 6 Year-by-year change in the effect of arrest on suspect homicide

	Number of suspect deaths		Risk ratio, 90 % CI			Exact tests (α=0.05)		
Year	Arrested	Warned	Estimate	Lower	Upper	One-sided	Two-sided	Cohen's *d*
1	1	0	.	.	.	0.6702	1.0000	
2	3	1	1.4762	0.2216	9.8347	0.5989	1.0000	−0.1703
3	5	1	2.4603	0.4072	14.8667	0.3576	0.6700	−0.3489
4	5	1	2.4603	0.4072	14.8667	0.3576	0.6700	−0.3489
5	6	1	2.9524	0.5011	17.3953	0.2691	0.4364	−0.4003
6	7	1	3.4444	0.5954	19.9274	0.2000	0.2836	−0.4390
7	9	1	4.4286	0.7846	24.9974	0.1072	0.1793	−0.4937
8	9	1	4.4286	0.7846	24.9974	0.1072	0.1793	−0.4937
9	10	2	2.4603	0.6911	8.7594	0.1866	0.3561	−0.3509
10	11	2	2.7063	0.7676	9.5421	0.1432	0.2404	−0.3789
11	11	3	1.8042	0.6212	5.2405	0.2689	0.5680	−0.2489
12	12	3	1.9683	0.6842	5.6620	0.2159	0.4090	−0.2799
13	12	3	1.9683	0.6842	5.6620	0.2159	0.4090	−0.2799
14	15	3	2.4603	0.8739	6.9265	0.1054	0.2052	−0.3529
15	15	3	2.4603	0.8739	6.9265	0.1054	0.2052	−0.3529
16	15	3	2.4603	0.8739	6.9265	0.1054	0.2052	−0.3529
17	15	3	2.4603	0.8739	6.9265	0.1054	0.2052	−0.3529
18	15	3	2.4603	0.8739	6.9265	0.1054	0.2052	−0.3529
19	15	3	2.4603	0.8739	6.9265	0.1054	0.2052	−0.3529
20	15	3	2.4603	0.8739	6.9265	0.1054	0.2052	−0.3529
21	15	3	2.4603	0.8739	6.9265	0.1054	0.2052	−0.3529
22	17	3	2.7884	1.0007	7.7696	0.0628	0.0963	−0.3898
23	17	3	2.7884	1.0007	7.7696	0.0628	0.0963	−0.3898

The risk of false discovery comes from running repeated tests on variables within the same, identical data set. That risk does not arise in a year-by-year analysis of this kind because each year constitutes a different data set. Our analysis does not consist of "mining" a sample to find one or more explanatory variables that survives after host of failed tests. Our analysis asks whether there is an emerging trend in a single relationship between two variables. We are not asking 22 different questions of the same data. Rather, we are asking one question of 22 different data sets. What we report is the answer to the question with which we began: how long does it take before any effect of the experimentally applied treatment becomes discernible? The answer can only be found by repeating the significance tests at constant intervals to reveal the precise date on which the risk of a chance result dropped below 10 %.

Whether the conventional significance level of 5 % is too conservative, or a marginal level of 10 % is too incautious, depends very much on a choice of philosophical perspectives (Ziliak and McCloskey 2008). In general, experimental statisticians prefer to consider confidence intervals, rather than a bright line of a 0.05 significance level. The confidence intervals for the homicide effect are clearly

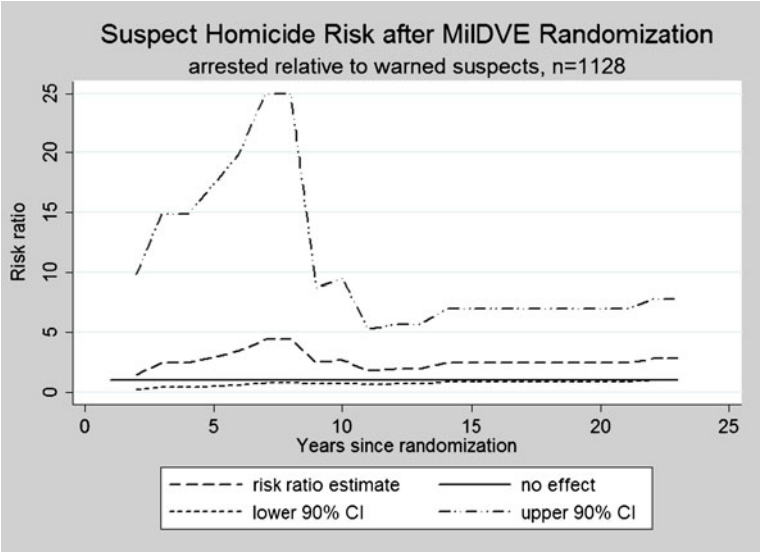

Fig. 4 Risk ratio plot showing risk of homicide for arrested versus warned suspects by year through year 23 post-randomization

wide, but they are shrinking over time. Whether they are still too wide for certain decisions is something we leave up to the readers.

Another way to analyze the 22 data points is observe that in all 22, the effect of arrest on homicide victimization is positive. Not once does the risk ratio go into reverse. Not once does the risk ratio go below a 47 % increase in homicides for those arrested over those warned. This constant signal amidst the noise of other causal influences is, perhaps, the most unlikely fact about the relationship between arrest and homicide. We cannot calculate a probability of this finding occurring by chance because the observations are not independent. However, we can say that the finding is consistent, over a long period of time (Fig. 4).

Homicide characteristics Qualitative evidence on the situational context of the suspects' homicide victimizations is sparse. In only 12 of the 20 homicides over the 23-year follow-up were we able to locate a homicide offense report, despite the substantial investment in the effort by the Milwaukee Police Department. Even when the reports were available, they only offered thin descriptions of circumstances in most cases. Yet most often, these descriptions fit the Wolfgang (1958) description of victim-precipitated homicide. Heavy drinking, arguments with friends, fights, and sudden rage predominate.

In one case, for example, the victim died of a stab wound inflicted by someone who confessed at the scene that he had just stabbed his "best friend" to death—with no further description of the events that had led to the stabbing. In another case, a man was drinking with two friends before he was beaten to death with a baseball bat. Events of this kind are described ethnographically by Anderson (1978) as encounters that may begin as playful, but may simply get out of hand (see also Collins 2008). Yet other homicides suggest miscalculations more than play: one offender was killed by two associates demanding

payment for drugs, while another was stabbed to death by his girlfriend when he refused to leave her home.

Exactly how and why these situations become more prevalent among the suspects randomly assigned to arrests many years earlier is not clear, but may be consistent with a non-linear effect of small differences. The lack of data on whether the victim-offender relationships in these murders was domestic or not is particularly unfortunate, since knowing that dimension would help us to specify the kind of conduct which might have provoked someone to kill them. From the evidence in the 12 available homicide reports, only three of the reports indicated the murder was committed by a domestic partner, all of which were in murders of decedents who had been randomly assigned to arrest. Even if there is a different rate at which suspects go on to cohabitate with partners between the arrested and warned suspects, it does not appear that the arrest-vs.-warn difference in homicide victimizations are explained by differences in domestic homicide.

Another way to examine these issues is to treat the research as a case–control study. Table 7 displays the prevalence of certain risk factors among both murdered and unmurdered suspects in the experiment. This shows that while arrest is not the leading post-dictive factor in homicide victimization, it is in the same range of risk elevation as such consistent predictors as race and prior arrest. The latter, however, both meet conventional significance levels, while arrest remains marginal. The difference between arrest and these other factors, of course, is that arrest is the only variable that can be manipulated by policy.

What about SUTVA (stable unit transfer value assumption)?

Any interpretation of causality in these findings, of course, depends upon whether they meet the SUTVA test (Rubin 1980). The central idea of SUTVA is that the treatment of one person should not affect the treatment of another person—or even the way each person perceives the treatment. Some treatments are far more likely than others to have "spillover" that spreads treatment effects into control conditions. In case-by-case arrest decisions occurring in residences, there is little direct observation of police decision-making by other citizens not affected by the case. In neighborhoods with a high

Table 7 Suspects at baseline who were not murdered versus those murdered

	Not murdered		Murdered		*t* test (α=0.05)	*p* value	Risk ratio
	n	%	*n*	%			
Arrested	739	66.70	17	85.00	−1.7263	0.08	1.27
Male	1,000	90.25	20	100.00	1.4684	0.14	1.11
Black	832	75.09	19	95.00	2.0522	0.04	1.27
Had a job	488	44.65	6	30.00	1.3064	0.19	0.67
Partner had a job	321	29.63	9	45.00	−1.5530	0.12	1.52
Had a prior arrest	704	63.54	17	85.00	−1.9825	0.05	1.34
Had ever been married	343	31.01	3	15.00	1.5385	0.12	0.48
	Mean	SD	Mean	SD			
Age	31.62	8.78	29.9	8.71	0.8679	0.39	0.95

prevalence of arrests for all categories of offenses, even word of mouth can entail great noise about what arrests occur for what causes. The long history of police discretion in not arresting for minor domestic assault, but arresting on some occasions (Reiss 1971; Black 1980), created the context in which this experiment was conducted. Save for those suspects who were encountered more than twice in the Milwaukee experiment, there is little basis for concern about whether the SUTVA assumption was violated.

Of those 72 suspects who were enrolled more than once in the Milwaukee experiment, only two of them were homicide victims. Both of them were arrested on the initial enrolment, and both were arrested on the repeat enrolment. To the extent that other suspects may have been encountered on prior or subsequent occasions by Milwaukee police, and arrested or warned for domestic violence, there may be many opportunities for mixed perceptions of sanction certainty or fairness by these suspects. Yet that "noise" is not the same as a SUTVA violation. It is true for all suspects in the sample, and throughout the entire 23-year follow-up period. Even the repeat encounters within the experiment fall into that general category of background context. What is important to note is that of the 20 murder victims, 18 of them were randomly assigned only once in the experiment. There was therefore no spillover of assigned treatments within those individuals. Whether there was spillover or conflict from encounters outside the experiment is arguably not relevant, since it was occurring in both treatment groups before and after enrolment.

Discussion

Our 23-year examination of mortality associated with arrest of domestic misdemeanor assault suspects shows no discernible differences in death rates from any cause of death except homicide. Whether that difference is discernible enough is a conclusion we leave to our readers. Given the magnitude of the effect size in the point estimate, even a marginally significant cause of homicide observed in a controlled experiment seems important to note in a policy debate affecting millions of arrests a year world-wide. That point estimate is substantively important, and consistent in direction over the entire follow-up period. The findings falsify the biological theory that arrest causes post-traumatic illnesses among the suspects, such as heart disease, that lead to earlier death, and supports a criminological theory of frustration and defiance leading to more incidence of victim-precipitated homicide. The findings are consistent, to the limited extent of the evidence, with a further influence of such stakes in conformity as employment and marriage, and an added strain of having the victim empowered by employment at the time of the arrest. Yet there is no indication that the suspects are generally killed by intimate partners, with only three out of 12 of the available cases suggesting a domestic argument as the precipitating factor.

The best theoretical explanation of these findings may come from Agnew's General Strain Theory (GST), which emphasizes the accumulation of frustrations and humiliations of daily life in predicting law breaking and aggression. It is precisely that prediction that would account for the suspects being murdered at a higher rate after they provoked others in their communities at a higher rate. The causal mechanism of this theory is not directly testable in our present data, but can be explored in subsequent analyses of differences in offending trajectories between murder victims and other suspects. It is also noteworthy that the effect of arrest on future homicide was almost entirely limited to suspects who had prior arrest records. That moderator effect seems

consistent with the GST concept of cumulative frustration, making suspects less resilient to withstand the effects of additional arrests.

The role of stakes in conformity first suggested by Jackson Toby (1957) may be more directly relevant to a theory of homicide victimization than Sherman's (1993) defiance theory that built on Toby's framework. Just as "stakes" may give a rational actor more to lose from delinquency—and thus exercise a protective effect against deviance—they may also exercise a protective effect against provocative behavior with peers that could precipitate a homicide. This theoretical idea must be treated with caution, however, since no interactions with measures of stake in conformity are significant as of 23 years; they only increase the effect size of arrest in a multivariate analysis. Perhaps a 33-year follow-up may provide clearer evidence on stakes in conformity and homicide victimization.

Other theories may explain these findings better than the frameworks we have considered. The idea of a public health science of criminal justice actions may expand into new domains, with these findings providing a small piece of the puzzle. Biologically mediated effects of arrest on mortality of domestic assault victims, as we report in another paper, will provide another piece. In general, however, we consider the causal mechanisms for the homicide causation we have identified to still be largely unexplained and largely unmeasured.

The policy findings are potentially far more direct at this point than any theoretical implications. At the least, our evidence justifies a fast track for replications of our test of homicide victimization by randomly assigned treatment. Just as the original Minneapolis experiment needed (and received) replication to extend our knowledge of the effects of arrest, these findings require replication in order to assess their external validity. *Until findings from other experiments can be reported*, *the generalizability of our findings remains highly uncertain*.

The external validity of these findings is the most pressing question for the policy implications of this paper. From a UK perspective, for example, the first question would be whether the findings depend on the high level of gun ownership in US cities. If the findings depended on ready accessibility to guns, they could have had limited applicability to low gun-density societies. Yet because 11 of the 20 homicides—over half—were committed by means other than gunfire, even low-gun societies could find that mandatory arrest for domestic violence increases the overall rates of intentionally inflicted death and harm in their communities. The only way such societies can be sure that a mandatory arrest policy for minor domestic violence does not have lethal consequences is to conduct their own replications of the MilDVE. Since no such experiment has ever been allowed in the UK, the most immediate point about external validity is the need to conduct a randomized test of mandatory arrest. This is especially true since the vast majority of arrests for domestic abuse in the UK result in suspects being released from a police station without being charged, just as they were in the Milwaukee experiment.

The key point about all such experiments is that they are not based on the serious tip of the iceberg of domestic violence, but on the majority of police encounters with domestic violence in which little physical injury has occurred. The issue at hand is not arrest for committing serious medical harm, in most cases. Such cases were screened out by design. The applicability of the conclusions will always be limited to the high volume of cases falling below a very low threshold of seriousness.

Those eligibility conditions applied to all six of the domestic violence arrest experiments funded by the National Institute of Justice. The feasibility of replication

of these findings now depends entirely on whether the original investigators have retained the identifying details of the victims and suspects in the experiments. It appears, for example, that there are no surviving records of the Minneapolis experiment's victims and offenders, at least not in the possession of the principal investigator. Whether the PIs of the other four experiments can help locate identifiers for their studies is a question we have not yet explored. If they have, then the National Institute of Justice could increase its return on the original investment in these experiments by supporting one or more replications of our study of death rates among both victims and offenders in relation to arrest.

Conclusions

On the larger questions of how to pursue a comprehensive cost–benefit analysis of mandatory arrest policy, the challenge remains how to integrate findings at the individual and macro levels. It is entirely possible, for example, that mandatory arrest may have beneficial macro-level effects even while it has harmful effects at the individual level. It may even have different effects on different subgroups within experiments. Our observation that almost all of the murder victims had a prior arrest, for example, may mean that our conclusions are largely limited to that group. Yet there is no legal requirement that all misdemeanors be treated with a one-size fits all policy. As further evidence on the macro and micro-level effects of arrest is reported, and as a more robust theoretical interpretation of that evidence may be developed (Sampson et al. forthcoming), we may find clearer guidance from the research related to this policy debate.

Perhaps our firmest conclusion from our study at this point is that 2 years is not enough to assess the effects of arrest on individuals. The expected follow-up period for all crime prevention experiments should probably be discussed in decades, not years. The effects of policy decisions applied in each individual case may be long-lasting, and take a full life-course to emerge. If both policymakers and research funders adopted a longer-term perspective on what constitutes a complete impact evaluation, research would become more helpful as a guide to policy.

Acknowledgments We would like to thank Milwaukee Chief of Police Edward Flynn for his strong support of our data collection efforts; the late Police Chief Robert Ziarnik for his support of the original experiment; and David Mazeika and Brad Bartholomew for excellent research assistance. We also offer our appreciation for the excellent service provided to the followup data collection by Joyce Knapton at Wisconsin's Department of Health Services and Dan Polans and Carianne Yerkes of the Milwaukee Police Department.

References

Agnew, R. (2006). *Pressured Into Crime: An Overview of General Strain Theory.* Los Angeles: Roxbury Publishing.

Anderson, E. (1978). *A Place on the Corner.* Chicago: University of Chicago Press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society.* Series B (Methodological), 289–300.

Berk, R. A., Campbell, A., Klap, R., & Western, B. (1992). The deterrent effect of arrest in incidents of domestic violence: A Bayesian analysis of four field experiments. *American Sociological Review, 57*(5), 698–708.
Bernard, T. J. (1990). Angry aggression among the "truly disadvantaged". *Criminology, 28*(1), 73–96.
Bushway, S. D. (1998). The impact of an arrest on the job stability of young white American men. *Journal of Research in Crime and Delinquency, 35*(4), 454–479.
Collins, R. C. (2008). *Violence: A Micro-Sociological Theory.* Princeton, NJ: Princeton University Press.
Dugan, L., Nagin, D. S., & Rosenfeld, R. (2003). Exposure reduction or retaliation? The effects of domestic violence resources on intimate-partner homicide. *Law & Society Review, 37*(1), 169–198.
Gleick, J. (1987). *Chaos theory: Making a new science*. New York: Viking.
Iyengar, R. (2009). Does the certainty of arrest reduce domestic violence? Evidence from mandatory and recommended arrest laws. *Journal of Public Economics, 93*(1–2), 85–98.
Jacobs, J. A. (2007). Further Reflections on ASR's Greatest Hits. *The American Sociologist, 38*(1), 99–131.
Kubzansky, L. D., Koenen, K. C., Spiro, A., III, Vokonas, P. S., & Sparrow, D. (2007). Prospective study of posttraumatic stress disorder symptoms and coronary heart disease in the Normative Aging Study. *Archives of General Psychiatry, 64*(1), 109.
Lauritsen, J. L., Sampson, R. J., & Laub, J. H. (1991). The link between offending and victimization among adolescents. *Criminology, 29*(2), 265–292.
Lempert, R. (1989). Humility is a virtue: On the publicization of policy-relevant research. *Law and Society Review, 23*(1), 145–161.
Massey, D. S. & Denton, N. A. (1993). American apartheid: Segregation and the making of the underclass. Cambridge, MA: Harvard University Press.
McCord, J. (1978). A thirty-year follow-up of treatment effects. *American Psychologist, 33*(3), 284.
McCord, J. (1981). Consideration of some effects of a counseling program. In S. Miller (Ed.) *New directions in the rehabilitation of criminal offenders* (p. 394–405). Washington, DC: National Academies Press.
Maxwell, C. D., Garner, J. H., & Fagan, J. A. (2002). The preventive effects of arrest on intimate partner violence: research, policy, and theory. *Criminology & Public Policy, 2*(1), 51–80.
Nores, M., Belfield, C. R., Barnett, W. S., & Schweinhart, L. (2005). Updating the economic impacts of the High/Scope Perry Preschool program. *Educational Evaluation and Policy Analysis, 27*(3), 245–261.
Oquendo, M., Friend, J. M., Halberstam, B., Brodsky, B. S., Burke, A. K., Grunebaum, M., Malone, K. M., & Man, J. J. (2003). Association of comorbid posttraumatic stress disorder and major depression with greater risk for suicidal behavior. *The American Journal of Psychiatry, 160*(3), 580–582.
Pate, A. M., & Hamilton, E. E. (1992). Formal and informal deterrents to domestic violence: The Dade County Spouse Assault Experiment. *American Sociological Review, 691–697*.
Paternoster, R., Brame, R., Bachman, R., & Sherman, L. W. (1997). Do fair procedures matter? The effect of procedural justice on spouse assault. *Law and Society Review*, 163–204
Rubin, D. B. (1980). Comment on "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *Journal of the American Statistical Association, 75*, 591–93.
Sampson, R. J., Winship, C., & Knight, C. (Forthcoming). Translating Causal Claims: Principles and Strategies for Policy-Relevant Criminology. *Criminology and Public Policy.*
Sherman, L. W. (1984). Experiments in Police Discretion: Scientific Boon or Dangerous Knowledge? *Law and Contemporary Problems, 47*(4), 61–81.
Sherman, L. (1992). *Policing domestic violence: Experiments and dilemmas*. New York: Free Press.
Sherman, L. W. (1993). Defiance, deterrence, and irrelevance: A theory of the criminal sanction. *Journal of Research in Crime and Delinquency, 30*(4), 445–473.
Sherman, L. W., & Berk, R. A. (1984). The specific deterrent effects of arrest for domestic assault. *American Sociological Review, 49*(2), 261–272.
Sherman, L. W., & Cohn, E. G. (1989). The impact of research on legal policy: the Minneapolis Domestic Violence Experiment. *Law and Society Review, 23*(1), 117–144.
Sherman, L. W., Harris, H. M. (2013a). *Effects of arrest over the life-course: A 24-year follow-up of the Milwaukee domestic violence experiment.* Paper presented at the Stockholm Criminology Symposium, June 11th, Stockholm, Sweden.
Sherman, L. W., Harris, H. M. (2013b). *Increased mortality of domestic assault victims after arrest of their partners: 23–year evidence from a randomised, controlled trial*. Under Review.
Sherman, L. W., Schmidt, J. D., Rogan, D. P., Gartin, P. R., Cohn, E. G., Collins, D. J., & Bacich, A. R. (1991). From initial deterrence to long-term escalation: short–custody arrest for poverty domestic violence. *Criminology, 29*(4), 821–850.
Sherman, L. W., Schmidt, J. D., Rogan, D. P., & Smith, D. A. (1992). Variable effects of arrest on criminal careers: The Minneapolis Domestic Violence Experiment. *Journal of Criminal Law & Criminology, 83*(1), 137.

Sherman, L. W. & Smith, D. A. (1992). Crime, punishment, and stake in conformity: legal and informal control of domestic violence. *American Sociological Review, 57*(5), 680–690.

Toby, J. (1957). Social disorganization and stake in conformity: complementary factors in the predatory behavior of hoodlums. *The Journal of Criminal Law, Criminology, and Police Science, 48*(1), 12–17.

Wolfgang, M. (1958). *Patterns in Criminal Homicide*. Philadelphia: University of Pennsylvania Press.

Ziliak, S. T., & Mccloskey, D. N. (2008). *The cult of statistical significance: how the standard error costs us jobs, justice and lives*. Ann Arbor, MI: University of Michigan Press.

Lawrence W. Sherman is a Distinguished University Professor at the University of Maryland's Department of Criminology and Criminal Justice, as well as Wolfson Professor of Criminology and Director of the Jerry Lee Centre for Experimental Criminology at the Cambridge University Institute of Criminology. He designed and directed both the Minneapolis (MinDVE) and Milwaukee (MilDVE) domestic violence experiments in 1980–83 and 87–88, respectively.

Heather M. Harris is a doctoral candidate in Criminology and Criminal Justice at the University of Maryland in College Park. She has been a Summer Associate at the RAND Corporation and participated in the Bureau of Justice Statistics workshop in the Quantitative Analysis of Crime and Criminal Justice.