Estimating the Safety Impacts in Before-After Studies Using the Naïve Adjustment Method

Pei-Fen Kuo*
Assistant Professor
Department of Crime Prevention and Correction
Central Police University, Taiwan
Tel. (886) -033282321
Email: kpf@mail.cpu.edu.tw

Dominique Lord
Professor
Zachry Dept. of Civil Engineering
Texas A&M University, College Station
TX 77843-3136
Phone: (979) 458-3949
Fax: (979) 845-6481
Email: d-lord@tamu.edu

Submitted for potential publication in the
Journal of Transportmetrica A: Transport Science

May 31, 2017

*Corresponding Author

Estimating the Safety Impacts in Before-After Studies Using the Naïve Adjustment Method

Pei-Fen Kuo[1], Dominique Lord[2]

ABSTRACT

The before-after study is the most popular approach for estimating the safety impacts of an intervention or treatment. Recent research, however, has shown that the most common before-after approaches can still provide a biased estimate when an entry criterion is used and when the characteristics of the treatment and control groups are dissimilar. Recently, a new simple method, referred to as the Naïve Adjustment Method (NAM), has been proposed to mitigate the limitations identified above. Unfortunately, the effectiveness of the NAM using "real" data has not yet been properly investigated. Hence, this paper examined the accuracy of the NAM when the treatment group contains sites that have different mean values. Simulated and two observed datasets were used. The results show that the NAM outperforms the Naïve, the Control Group, and the empirical Bayesian methods.. Furthermore, it can be used as a simpler alternative for adjusting the Naïve estimators documented in previous studies.

Keywords: Before-after study, Naïve Adjustment method, Regression-to-the-mean, Safety effectiveness, Site selection bias

[1] Corresponding author, Crime Prevention and Corrections Department, The Central Police University, No.56, Shujen Rd., Takang Vil., Kueishan District, Taoyuan City 33304, Taiwan, Tel. 886-3-3282321, Fax: +886-3-3284118, Email: peifenkuo@gmail.com

[2] Zachry Department of Civil Engineering, Texas A&M University, College Station, TX 77843-3136, Phone: (979) 458-3949, Fax: (979) 845-6481, Email: d-lord@tamu.edu

INTRODUCTION

The before-after study is still the most popular approach for estimating the safety impacts of an intervention or potential treatment (Persaud and Lyon, 2007). Numerous studies have proposed various estimation methods for minimizing important biases, including the well-documented regression-to-the-mean (RTM) (Hauer, 1980 a,b, 1997; Abbess et al., 1981; Hauer et al., 1983; Danielsson, 1986; Wright et al., 1988; Rock,1995; Hamed et al., 1999; Davis, 2000; Miranda Moreno et al., 2009; Maher and Mountain, 2009; Lee et al, 2010; Shin et al, 2012; Tse et al 2014). Among the methods proposed to minimize or eliminate these biases, we find the empirical Bayes (EB) and the Control Group (CG) methods that are considered the most popular (Hauer, 1997). In particular, the EB method has become widespread since it can account for the RTM, and recent work has simplified its application for analyzing crash data (Persaud and Lyon, 2007; AASHTO, 2010). Despite this, recent studies have shown that the EB and CG methods could still suffer from important methodological limitations. For instance, Lord and Kuo (2012) and Kuo and Lord (2013) have revealed that the EB and CG methods provide a biased estimate when an entry criterion is used and when the characteristics of the treatment and control groups are dissimilar (e.g., different sample mean and variance values). In fact, in most cases, the key characteristics used for estimating the parameters associated with the treatment and control groups are either unknown or not known with certainty which makes this problem even more significant.

Recently, a new adjustment method (referred hereafter as the Naïve Adjustment Method or NAM) has been proposed for minimizing the problems identified above (Lord and Kuo, 2012; Kuo and Lord, 2013). This new method provides a more precise estimate than the Naïve approach and performs better than the CG and EB methods when control group data are not available or do not have the same characteristics. Moreover, the authors showed that the site-selection bias generated by using a dissimilar control group might be even higher with the CG and EB methods than with the Naïve method. Although full regression models or safety performance functions (SPFs) combined with the EB method can provide a better estimate than using flow-only SPF, the site selection bias will still exist if there is an entry criterion, no matter how many variables are used in the model. The issue is the characteristics of the truncated distribution, which is not dependent on the number of variables in the model, as discussed by Lord and Kuo (2012). This issues also applies to CM-Functions, as proposed by Chan and Persaud (2014).

For the bias to be completely eliminated, the distribution of every variable in the treatment and reference groups has to be exactly the same: mean, variance, skewness (which affects the dispersion), etc. It is true that CM-Functions better capture the characteristics of the treatment and reference groups, but in practice they are rarely the same. If they are, the sites included as part of the reference group should technically be considered as being candidate for treatment. Obviously, a complete randomize trial or analysis would remove the site selection effect or bias, but this is rarely if ever done in highway safety studies, as discussed by Hauer (1997).

So far, the analyses performed by Lord and Kuo (2012) have been theoretical in nature and used a simulation protocol in which the sample mean was assumed to be fixed (all sites have the same long-term mean) in order to accurately estimate different biases. In practice, however, sites used to collect datasets for the treatment and control groups have different long-term mean values, which can also influence the variance (see Lord, 2006). The long-term mean value, in this instance, is based on real observed crash data. However, it is seldom known in practice, since the long-term is in fact estimated (see Hauer, 1997). In most cases, the evaluation of treatments is based on data that were already collected, which means that the safety analyst does not know the "true" long-term mean for each site and cannot therefore control for external factors that can influence before-after studies.

Along the same line, the full Bayes (FB) models and propensity scoring methods (PSM) have recently been introduced for better estimating the safety of treatments using the before-after studies in highway safety (Park et al., 2010; Wood et al., 2015). However, it is important to point out that the selection effects will still influence the results of FB models if it is not specifically accounted for (via truncated models), as discussed in our previous study (Lord & Kuo, 2012). Wood & Donnell (2017) corroborated the theoretical results of Lord and Kuo (2012) using observed data. Furthermore, although Sacchi and Sayed (2015) did not specifically analyze site selection effects based on an entry criterion (they selected sites based on high estimated values), they compared several methods used for before and after studies, including the EB and FB methods, and found that the difference in bias in estimated values between the two methods were not statistically significant. Nonetheless, further work should be done for examining the how the site selection effects specifically influence FB models. Propensity scoring methods can also minimize for the selection bias, but the variables need to be specifically matched, which is very difficult to do in observational studies (see Winkelmayer & Kurth, 2004, Austin, 2011, Biondi-Zoccai et al., 2011, Lim et al., 2014 for important limitations associated with the PSM). As discussed in the paper, variables in the treatment and control groups can be very different, since the sites are selected after the treatments were implemented. The approach documented in this paper is in fact more straightforward and easier to implement.

Hence, the goals of this paper are therefore two-fold. The first objective was to examine the accuracy of the NAM when each site in the treatment group has a different (long-term) mean value. To accomplish this objective, both simulated and observed data were utilized. The second objective consisted in describing how the NAM can be easily used by researchers and transportation safety analysts or practitioners for evaluating the effectiveness of a treatment when the true mean of each site in the treatment group is not known and/or data collected for the control group are not available. An example to illustrate how the method can be used is presented.

BACKGROUND

Site-selection effect occurs when an entry criterion is used for selecting observations that will be included in the before-after study. For count data, this gives rise to a truncated negative binomial (NB) distribution. Figure 1 illustrates the characteristics associated with the site-selection effect for count data. In order to show the real magnitude of the RTM bias, we assumed that the treatment did not work, and then the treatment did not change the crash mean value in the before and after periods ($\Lambda_A = \Lambda_B$). If this treatment has no effect or is even, we can say that there is no real benefit in implementing the treatment.

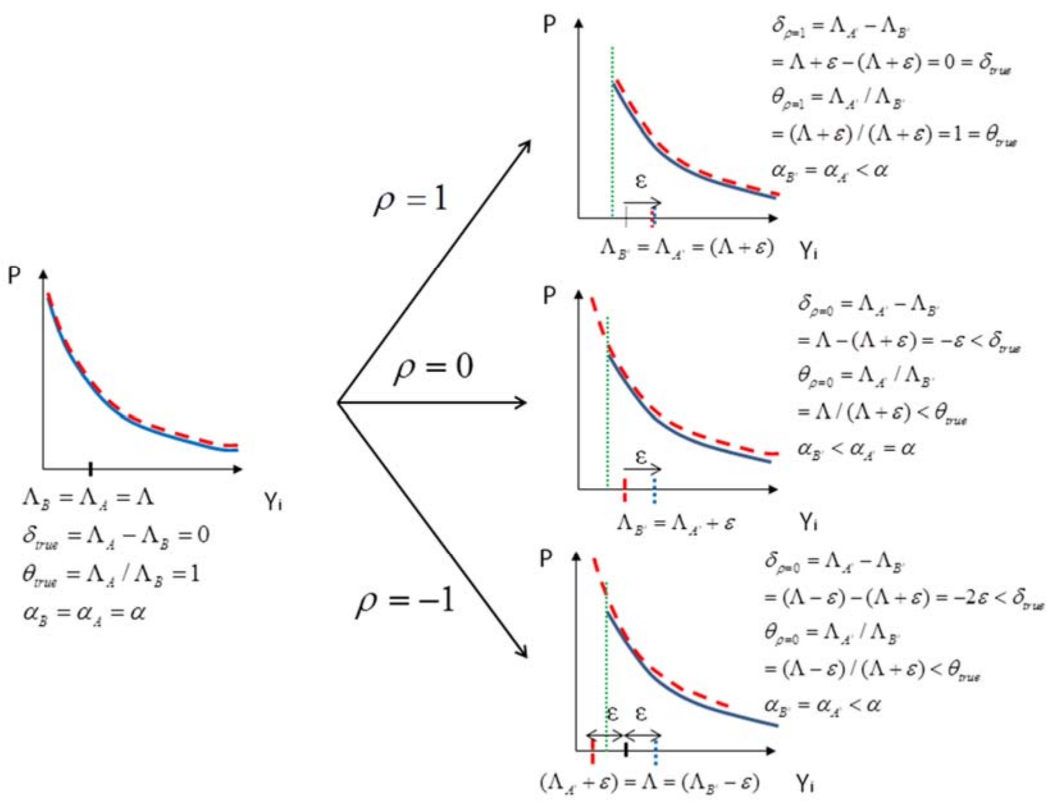

Fig. 1. Site-selection bias for different correlation coefficient values.

In Figure 1, the left-hand graph shows the probability related to the NB distributed data for the before and after periods (without selection). The blue solid line represents the before period, while the red dashed line is used to denote the after period. The difference between the two means and the ratio between the means (often referred to as *index of safety effectiveness*, see Hauer, 2007; for the rest of the paper, we refer to this index as the *safety index*) of these two curves are defined as δ_{true} and θ_{true}, respectively, and the dispersion parameters of these two curves are defined as α_B and α_A, respectively. In Figure 1, it is assumed that the applied treatment did not work, i.e., $\Lambda_A = \Lambda_B$, in order to show the selection biases of δ_{true} and θ_{true}. Those are displayed in the charts located on the right-hand side of Figure 1. After setting the minimum entry criterion (the green vertical dotted line), the data distribution in the before period is left-truncated, as indicated by the curved solid line in the right-hand side charts. The entry criterion is the threshold to select the site that is or will be included in the treatment group. For example, if our entry criterion is 5 crashes per year, and say an intersection with 5 or more crashes per year happened during the study period (similar to what is proposed in the Manual on Uniform Traffic Control Devices or MUTCD, FHWA, 2009), the site will be selected in the study; otherwise this site will not be included.

Comparing the three charts on the right-hand side, we found that setting the same entry criteria can cause different effects on the bias as a function of different correlation coefficient (ρ) values between the before and after data. If ρ is equal to 1, the estimator of the difference and the index ($\delta_{\rho=1}$ and $\theta_{\rho=1}$) are unbiased because the mean value increases in the same manner in the before and after periods. However, the estimator of the dispersion parameter ($\alpha_{\rho=1}$) is smaller because of the smaller variance.

If ρ is equal to 0, the Naïve estimator of the difference and the safety index ($\delta_{\rho=0}$ and $\theta_{\rho=0}$) are smaller than their true values. The mean in the before period increases, while the mean in the after period remains constant because the before and after data are independent. If ρ is negative, the estimator of the difference and the safety index ($\delta_{\rho=-1}$ and $\theta_{\rho=-1}$) become smaller than the above values. Removing data or observations with low values in the before period may also remove data with higher values in the after period because of the negative correlation. Hence, the difference between the mean values becomes smaller because of the higher mean in the before period and the lower mean in the after period.

Many researchers have assumed that the site-selection effect and the RTM are the same bias. In theory, they are not, as discussed in Davis (2000) and Lord and Kuo (2012). It should be noted that with site-selection effects, the entry criteria (C) can take any value and is limited not only to one criterion. On the other hand, with the RTM, the observations in the before period Y_{i1} are selected based only on the mean μ_1. In other words, the site-selection bias may still exist in some conditions where the RTM is ignored or not applicable, such as when the entry criterion is relative low (i.e., close to $Y_{ij} > 0$) or are used to remove extreme values by setting a maximum and minimum threshold for including observations in the dataset.

ESTIMATORS

There are four methods that can be used for estimating the safety effectiveness based on count data: the Naïve before-after study, before-after study with a control group, the EB method, and the NAM. For convenience purposes, the common notations are listed below:

θ: The safety index (ratio between the mean values),

δ: The difference between the mean values,

C : The entry criterion,

n : The sample size,

α : The dispersion parameters

$\Lambda_{i1}^{T}, \Lambda_{i1}^{C}$: The mean response rate for site i (T: treatment group, C: control group) in the *before* period, i=1, …, n,

$\Lambda_{i2}^{T}, \Lambda_{i2}^{C}$: The mean response rate for site i (T: treatment group, C: control group) in the *after* period,

N_{ij1}^{T}, N_{ij1}^{C}: The observed response for site i (T: treatment group, C: control group) in j year (in the *before* period) for count data, N_{ij1}^{T} > C,

N_{ij2}^{T}, N_{ij2}^{C} : The observed response for site i (T: treatment group, C: control group) in j year (in the *after* period) for count data. Here we let j=t=1 for calculation convenience purposes,

W_i: The weight for site i in the EB method, $W_i = \dfrac{1}{1+\hat{\Lambda}_{i1}^{T}\times\alpha}$,

$\hat{\Lambda}_1$: The estimator for the average crash rate of all sites in the *before* period,

$\hat{\Lambda}_{i1}^T = E(N_{i1} | N_{i1} > C) = \mu_1$ for the EB_{MM} method, and $\hat{\Lambda}_{i1}^T = \Lambda_{i1}^C$ for the EB_{CG} method.

M_{i1} : The expected responses for site i in the EB method,

$$M_{i1} = W_i \times (\Lambda_{i1}) + (1 - W_i) \times (\sum_{j=1}^{t} N_{ij1}).$$

In traffic safety studies, a treatment is considered effective if the crash mean in the after period is lower than in the before period, i.e. $\theta < 1$ (or $\delta < 0$). Given the notation above, it is now possible to define the equations for the safety index, the difference, and dispersion parameter, which are estimated using the four methods based on a truncated count distribution, as shown in Table 1 below. It is important to note that the *EB with MM* method was applied when no other reference group (control group) was available. With that method, the estimators of the crash mean and variance are based on the treatment group data only. The *EB with CG* method uses the same formula to calculate the empirical weight, but the estimators of crash mean and variance are based on the external control group data. The above methods were called "the simple moment method" and "the multiple variable method" in Hauer (1997, page 189). As for the CG method, it also includes control group data but was simply used as an adjusting ratio.

Table 1. Different types of safety indices by using common methods in before-after studies (Hauer, 1997; Persaud, 2001; Davis, 2000)

	$\hat{\theta}$ (Safety Index)	$\hat{\delta}$ (Difference)	$\hat{\alpha}$
Naïve method	$\frac{\hat{\Lambda}_2^T}{\hat{\Lambda}_1^T}=\frac{\frac{1}{n}\frac{1}{t}\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij2}^T}{\frac{1}{n}\frac{1}{t}\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij1}^T}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij2}^T}{\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij1}^T}$ (1)	$\hat{\Lambda}_2^T-\hat{\Lambda}_1^T=\frac{1}{n}\frac{1}{t}(\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij2}^T-\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij1}^T)$ (5)	$\frac{\left[\frac{E((N_{i2}-\Lambda_2)^2 \mid N_{i1}>C)}{E(N_{i2}\mid N_{i1}>C)}-1\right]}{E(N_{i2}\mid N_{i1}>C)}$ (9)
CG method	$\frac{\hat{\Lambda}_2^T}{\hat{\Lambda}_1^T\times\frac{\hat{\Lambda}_2^C}{\hat{\Lambda}_1^C}}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij2}^T}{\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij1}^T\times\frac{\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij2}^C}{\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij1}^C}}$ (2)	$\hat{\Lambda}_2^T-\hat{\Lambda}_1^T\times\frac{\hat{\Lambda}_2^C}{\hat{\Lambda}_1^C}=\frac{1}{nt}(\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij2}^T-\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij1}^T\times\frac{\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij2}^C}{\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij2}^C})$ (6)	
EB method	$\frac{\hat{\Lambda}_2^T}{\hat{\Lambda}_1^{EB}}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij2}^T}{\sum_{i=1}^{n}M_{i1}^T}=\frac{\Sigma_{i=1}^{n}\Sigma_{j=1}^{t}N_{ij2}^T}{\mathrm{W_i}\times(\Lambda_{\mathrm{i1}})+(1-\mathrm{W_i})\times(\sum_{\mathrm{j=1}}^{\mathrm{t}}N_{\mathrm{ij1}}^T)}$ (3)	$\hat{\Lambda}_2^T-\hat{\Lambda}_1^{EB}$ $=\frac{1}{n}\left\{\sum_{i=1}^{n}\sum_{j=1}^{t}N_{ij2}^T-\left[\sum_{i=1}^{n}\mathrm{W_i}\times(\Lambda_{\mathrm{i1}})+(1-\mathrm{W_i})\times(\sum_{\mathrm{j=1}}^{\mathrm{t}}N_{\mathrm{ij1}}^T)\right]\right\}$ (7)	
NAM	$\hat{\theta}_{naive}+\theta\left[\frac{\mu_1-\Lambda_1}{\mu_1(\Lambda_1\alpha+1)}\right]$ $=\hat{\theta}_{naive}\left[1+\frac{C+1 / (1+\frac{P(N>C+1)}{P(N=C+1)})}{\hat{\Lambda}_{1,naive}+\left[\frac{C+1}{(\hat{\Lambda}_{1,naive}\hat{\alpha}_{naive}+1)^{-1}(1+\frac{P(N>C+1)}{P(N=C+1)})}\right]}\right]$ (4)	$\hat{\delta}_{naive}-(\mu_1-\Lambda_1)\frac{(\delta\alpha-1)}{(\Lambda_1\alpha+1)}=\hat{\delta}_{naive}-\left[\frac{(C+1)\times(\hat{\delta}_{naive}\hat{\alpha}_{naive}-1)}{1+\frac{P(N>C+1)}{P(N=C+1)}}\right]$ (8)	$\hat{\alpha}_{naive}+\hat{\alpha}_{naive}^2\left[\frac{\nu_1-\mu_1-\mu_1^2\hat{\alpha}_{naive}}{(\mu_1\hat{\alpha}_{naive}+1)^2}\right]$ (10)

It should be noted that Equation (9) in Table 1 was calculated using crash data in the after period instead of using crash data in the before period, a technique that is commonly used in current before-after studies, i.e.

$$\hat{\alpha}_{naive} = \frac{\left[\frac{E((N_{i2}-\Lambda_2)^2 \left| N_{i1} > C\right.)}{E(N_{i2} \left| N_{i1} > C\right.)} - 1\right]}{E(N_{i2} \left| N_{i1} > C\right.)}$$

$$\text{instead of } \hat{\alpha}_{naive} = \frac{\left[\frac{E((N_{i1}-\Lambda_1)^2 \left| N_{i1} > C\right.)}{E(N_{i1} \left| N_{i1} > C\right.)} - 1\right]}{E(N_{i1} \left| N_{i1} > C\right.)} \quad (11)$$

SIMULATED DATASET: VARYING CRASH MEAN VALUES

This section describes the simulation analysis carried out to examine the accuracy of the NAM when each site in the treatment group has a different mean value. The first part describes the simulation protocol, while the second part covers the simulation results.

Simulation Protocol

According to Lord and Kuo (2012), Equations (4), (8), and (10) can partially remove site-selection bias even when similar control group data are not available. These equations could potentially remove up to about half the bias if the crash mean is assumed to be fixed (i.e., all sites have the same mean) and if the crash counts are assumed to follow a NB distribution. As discussed above, the assumption that each site has the same long-term mean is not realistic, since sites included in the sample are geographically located far apart from each other or have different characteristics, such as lane configurations (not captured in the data collection process for example) and driver/vehicle compositions among others. Even if the sites have the same traffic flow volume (assuming that it is correctly estimated), they are expected to have different long-term mean values or estimates (see Lord et al., 2005).

To evaluate the robustness of the equations described in the previous section (only those associated with the safety index), the authors updated the modeling protocol proposed by Lord and Kuo (2012). Here, the population of the crash mean, Λ_1, is assumed to follow a lognormal distribution, similar to the procedure described in Lord (2006). This assigns a different crash risk to each observation, while still maintaining that the crash count for each site follows a NB distribution (i.e., Poisson-gamma).

In the simulation, the number of crashes followed a mixture of the Poisson-gamma and log-normal distributions. The variance σ of the lognormal distribution was changed as follows: 0 (fixed crash mean), 0.01 (small heterogeneity), 0.5 (median heterogeneity), and 1 (large heterogeneity). The other input variables, such as the sample mean value and dispersion parameter, were kept the same as those described in Lord and Kuo (2012). The simulation protocol was as follows:

$$\begin{aligned} &N_{ik} \sim Poisson(u_i \Lambda_k) \\ &\Lambda_1 \sim \log normal(\log(3), \sigma) \\ &u_i \sim Gamma(\alpha^{-1}, \alpha) \end{aligned} \tag{11}$$

Where,

N_{ik} : The crash frequency for site i in the k period,

u_i : The subject-specific random effect,

Λ_1 : The crash mean in the before period,

σ : The variance of the crash mean,

α : The dispersion parameters.

We used the dispersion parameters and the variance of the crash means to indirectly characterize the spatial effects. As described above, if two 'similar' sites follow a distribution with same crash mean values, it is expected that the observed crash count for each site will be different because of the dispersion parameter or the variation in the data. If the distribution has a large variance, then the counts between the two sites should be even greater because of the spatial effects (note: spatial regression models reduce the variation or variance observed in the data). In the same situation, the crash frequencies for different years for the same site might be still be different because the correlation coefficient values between the before and after periods may not be equal to one.

Simulation Results

This section describes the results of the simulation analysis. Figure 2 shows the simulation results for different variances using four before-after methods [six scenarios: the Naïve, EB (method of moment/control group), CG (similar/non-similar control group), and NAM]. The entry criterion is five crashes per year, and the crash mean of the non-similar control group is five times more than that of the treatment group. Figure 2 (a) to (d) clearly shows a trend that larger variances in the mean distributions decrease site-selection biases. This was expected because larger variances among the crash means increase the between-subject variance (similar to the effects of larger dispersion parameter values), resulting in lower site-selection bias (see Hauer, 1997; Lord and Kuo, 2012). It should be noted that the NAM still reduced the site-selection bias by approximately 50% even in cases of very large heterogeneity. However, the estimator might overestimate the site-selection bias when the entry criteria are low (Figure 3). Compared to all methods, the CG method (based on a perfect control group) provides a more precise estimate than the NAM, but the latter is much better than the EB_{CG}, EB_{MM}, non-similar (NS) CG, and the Naive methods.

Although simulated data may not completely reproduce the 'real effectiveness' of a treatment (such as the installation of red-light-running cameras or widening of the road width) because one cannot account for the unknown factors that influence crash risk, it is necessary for calculating biases of statistical models or methods. Real or observed data do not allow for the identification or calculation of biases or systemic errors, since the true values for the mean, variance, and skewness, etc. are not known with certainty. They are in fact estimated. Simulation is frequently used by statisticians for evaluating the characteristics of models and mathematical methods.

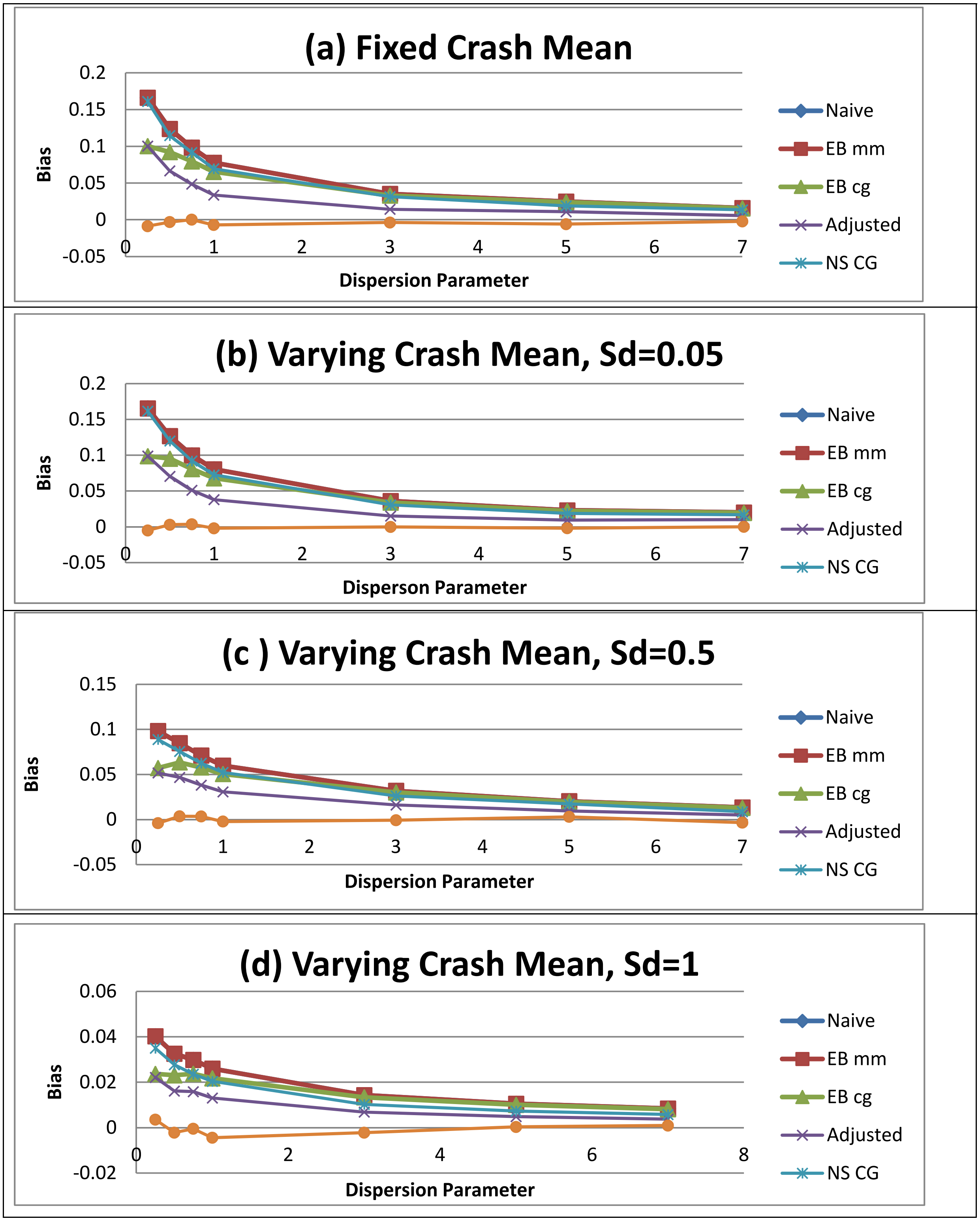

Fig. 2. Site-selection biases for different before-after methods when the standard deviation of the crash mean is equal to 0, 0.01, 0.5, and 1

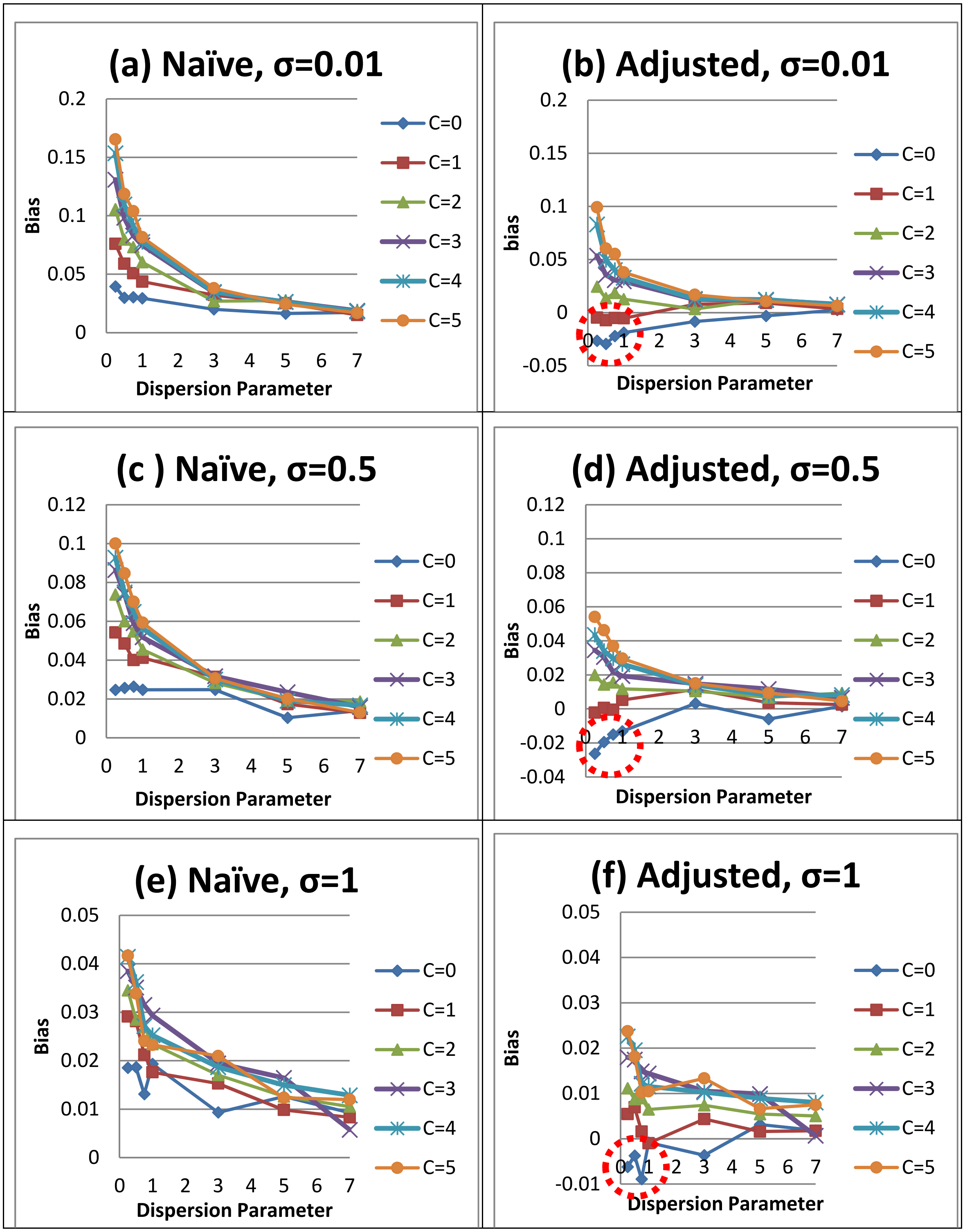

Fig. 3. Site-selection biases for Naïve and Adjustment Methods when the standard deviation of the crash mean is equal to 0.01, 0.5, and 1

OBSERVED COUNT DATASETS: APPLICATION OF THE ADJUSTMENT METHOD

In the previous section, the simulation results showed that the NAM reduced the site-selection bias by approximately 50% even when all sites have different crash mean values. In this section, the authors applied the NAM to two observed datasets in order to evaluate its practical effectiveness. The first dataset was collected in College Station, Texas. A dummy variable was used to represent a hypothetical treatment that was implemented in that city. The second dataset was assembled to evaluate the effects of red-light cameras on the overall number of crashes at signalized intersections. Also, only the Naïve and NAM were applied to the data because no control group data were available and no SPFs existed to apply the EB method (i.e., EB_{CG}).

Application #1: The Dummy Treatment Trial in College Station, Texas

Application #1 Study Design

The authors first assumed that there was a dummy treatment applied to 917 sites (which had at least one recorded crash in 2008) in College Station, Texas, on December 31, 2008. Therefore, the year 2008 was defined as the *before* period, while the year 2009 was defined as the *after* period. Because there was no such applied treatment, its safety effectiveness should be close to 1. Reader may refer Sacchi and Sayed (2015) for other method to improve the accuracy of No-treatment before-after evaluations. The actual safety index of this dummy treatment (based on all of the crash data available, without a positive criterion selection) was 0.95 or a 5% reduction. Then, the crash data were filtered by different entry criteria ranging from 1 to 5 crashes per year. The above procedure is used to mimic the process when traffic engineers set different entry criteria to select hot spots, apply a dummy treatment, and conduct a before-after crash study.

Step-By-Step Procedure for Using the Naïve Adjustment Method in Application #1

This section describes the step-by-step procedure for using the NAM in order to better illustrate how it can be used with observed data or to adjust the results documented in previous studies (if enough information is available). For this description, the entry criterion was set to 5 crashes per year, $Y > C = 4$. Among all sites, 121 sites were identified as having a crash record equal to or above 5 crashes per year in the before period: $Y_1^T = [5, 5, ..., 31, 32]$. The corresponding crashes in the after period are: $Y_2^T = [5, 4, ..., 41, 32]$. The authors have also created an Excel spreadsheet for anyone who wants to use it. This NAM calculation spreadsheet (see Appendix) can help readers to calculate an NAM value when the naïve estimator and C value (as an entry criterion) is available.

The steps are as follows:

Step 1: Calculate the naïve estimate.

$$\theta_{naive} = \frac{\Lambda_2}{\Lambda_1} = \frac{(5+5+...+31+32)}{(5+4+...+41+32)} = \frac{983}{1217} = 0.808$$

Step 2: Estimate the value of variables in the estimator of the Naïve Adjustment method (Equation 4).

$$\hat{\Lambda}_{1,naive} = \frac{1217}{121} = 10.06$$

$$\alpha_{naive=} \frac{\left[\frac{E((N_{i2}-\Lambda_2)^2 \,|N_{i1} > C)}{E(N_{i2}\,|N_{i1} > C)} - 1\right]}{E(N_{i2}\,|N_{i1} > C)} = \frac{(60.8/8.12)-1}{8.12} = 0.798$$

$\frac{P(N > C+1)}{P(N = C+1)} = \frac{0.600}{0.060} = 9.97$ (The probability and cumulative probability were estimated using the "dnbinom" and "dnbinom" functions in the R software program.)

Step 3: Calculate the safety effectiveness based on the Naïve Adjustment method (Equation 4).

$$\hat{\theta}_{naive} + \theta\left[\frac{\mu_1 - \Lambda_1}{\mu_1(\Lambda_1\alpha+1)}\right]$$

$$= \hat{\theta}_{naive}\left[1 + \frac{C+1 \Big/ (1+\frac{P(N>C+1)}{P(N=C+1)})}{\hat{\Lambda}_{1,naive} + \left[\frac{C+1}{(\hat{\Lambda}_{1,naive}\hat{\alpha}_{naive}+1)^{-1}(1+\frac{P(N>C+1)}{P(N=C+1)})}\right]}\right]$$

$$= = 0.808\left[1 + \frac{5/(1+9.97)}{10.06 + \left(\frac{5}{(10.06 \times 0.798+1)^{-1}(1+9.97)}\right)}\right]$$

$$= 0.834$$

Repeat the above steps to get the naïve and the adjusted estimators of the safety index for different entry criteria.

Application #1 Study Results

By following the same procedure as described in the previous section, we can get the results illustrated in Figure 4, which shows the Naïve and NAM estimators for the safety index (θ), differences in mean values (δ), and dispersion parameters (α) for different entry criteria (estimated by Equations (1), (4), (5), (8),(9) and (10)). The results clearly show that a higher entry criterion results in an overestimation of a treatment's safety effectiveness, especially when the Naïve method is used. Using the NAM could partially reduce the site-selection bias, which leads to the estimators being closer to the true value (0.95). In Figure 4(a), the estimators for the safety index (indicated by a green triangle line) are closer to the true value (illustrated by a blue diamond line) than when using the Naïve method (shown by a red square line), except when the entry criteria are very small. For the difference in sample means and dispersion parameters, the results were analogous to those of the safety index. Setting a higher entry criterion causes a higher level of site-selection bias. By partially removing site-selection bias, the NAM estimators were usually closer to the true value than for the Naïve estimators (Figure 4(b) and Figure 4(c)). Also, it should be noted that the distribution of this dataset followed a negative binomial distribution. For the crash data in the *before* and *after* periods, the inverse dispersion parameters are 1.694 (σ=0.113) and 0.587 (σ=0.037), and the mean parameters are 2.665 (σ=0.086) and 2.530 (σ=0.121), respectively. If the dataset does not follow the assumed distribution (negative binomial), equations (4), (8), and (10) would not be able to adjust site-selection biases properly.

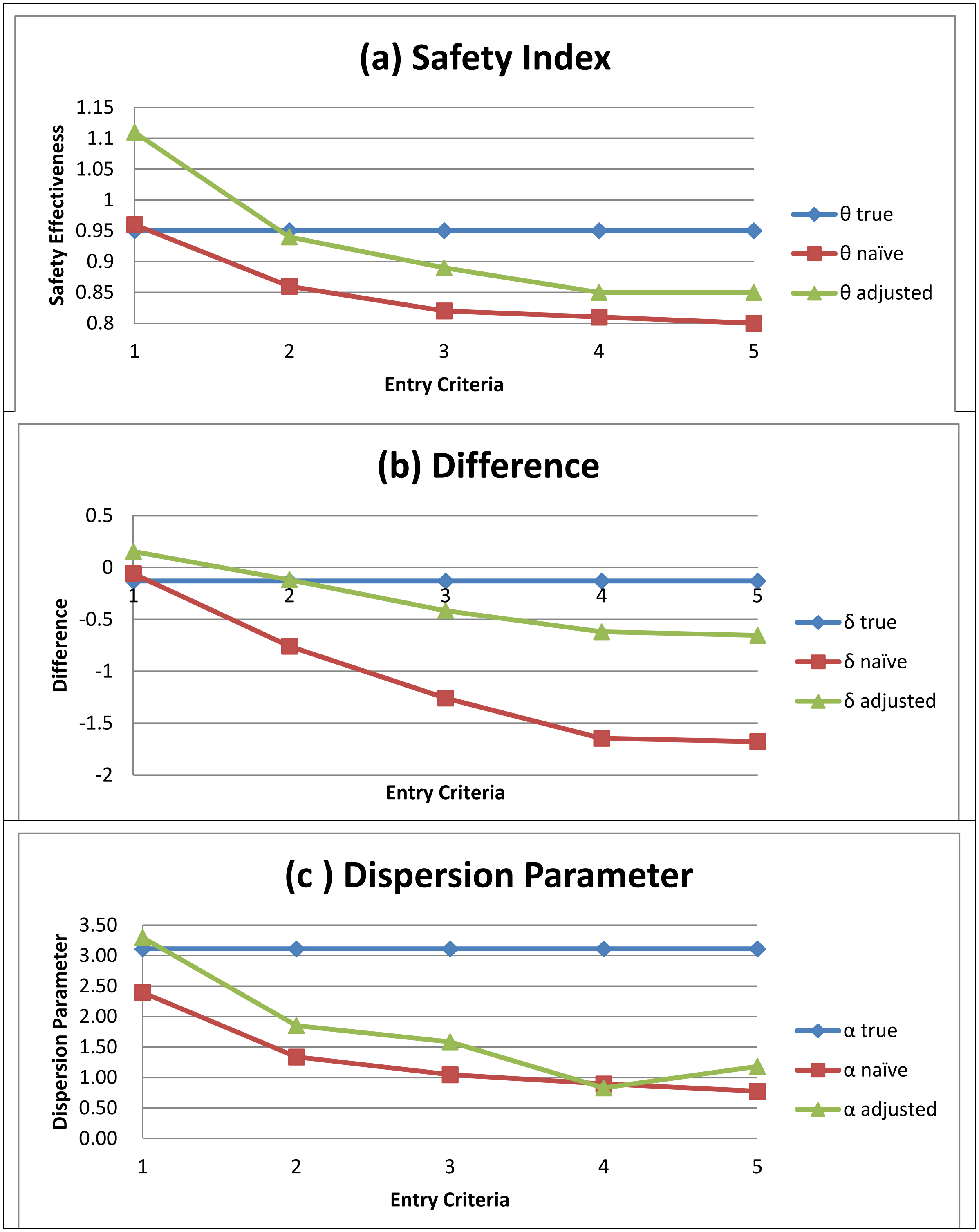

Fig. 4. Estimated safety effectiveness, difference, and dispersion parameter for the Naïve and Adjustment Methods

Application #2: The Red-Light Running Camera Trial.

Application #2 Study Design

The second dataset was obtained from a Texas A&M Transportation Institute (TTI) project, which examined the safety impacts of red-light running cameras on crash frequency. The original dataset included 319 signalized intersections located in different cities in Texas; of these, the authors selected 95 intersections, all of which had the same *before* and *after* periods (two years) in order to simplify the analysis. The crash mean value was 8.56 and the variance was 35.7. The minimum and maximum crash frequencies were 5 and 37. This dataset was ideal for examining the NAM because it offered a large sample size. Furthermore, some sites exhibited particularly low crash frequencies (even zero) in the before period. In other words, the estimators of the mean values, dispersion parameters, and the safety index could be treated as true values because there were no entry criteria for this dataset (because of the sites with zero crash), but with the important assumption that the dispersion parameter also represented the true value. Also, the distribution for the various samples was assumed to follow a NB distribution. It was presumed that the true values of θ, δ, α were 0.72, -1.93, and 0.77, respectively, since they were estimated from the whole population. The results obtained using the Naïve before-after method showed that the overall crash frequency for all intersections decreased by 6.8%. To examine the effects of the site-selection bias, Equation (4) was used to obtain the naïve adjusted safety index. The first entry criterion was set to 0, because the suggested initial assumption for the entry criterion should be considered just one unit below the smallest set of observed data (C=min N_{ij} -1) (Johnson et al., 1970). Entry criteria equal to 1, 2, 3, … up to 20 were subsequently employed.

Application #2 Study Results

Figure 5(a) shows the differences between the $\theta_{adjusted}$, θ_{naive}, and the θ_{true}. This figure clearly illustrates that the NAM estimators yielded a lower site-selection bias than the Naïve estimators. Moreover, Figure 5(a) shows that higher entry criteria tend to cause higher selection biases, which lead to an overestimation of safety index. Equation (4) partially eliminates the selection bias, but does not remove all the selection bias (as discussed above). Generally, Figure 5(a) supports the theoretical and simulation results described above, such that higher entry criteria cause larger (more negative) biases, and the NAM estimators are closer to the true value than those produced by the Naïve estimator. In addition, Figure 5 shows that higher entry criteria may lead to underestimate the dispersion parameter and overestimate of the difference in mean values. This outcome is consistent with the simulation results and the first observed dataset. However, it should be noted that low sample size may affect the effectiveness of the Adjusted method, especially when entry criteria are over 15 (Figure 5 (d)).

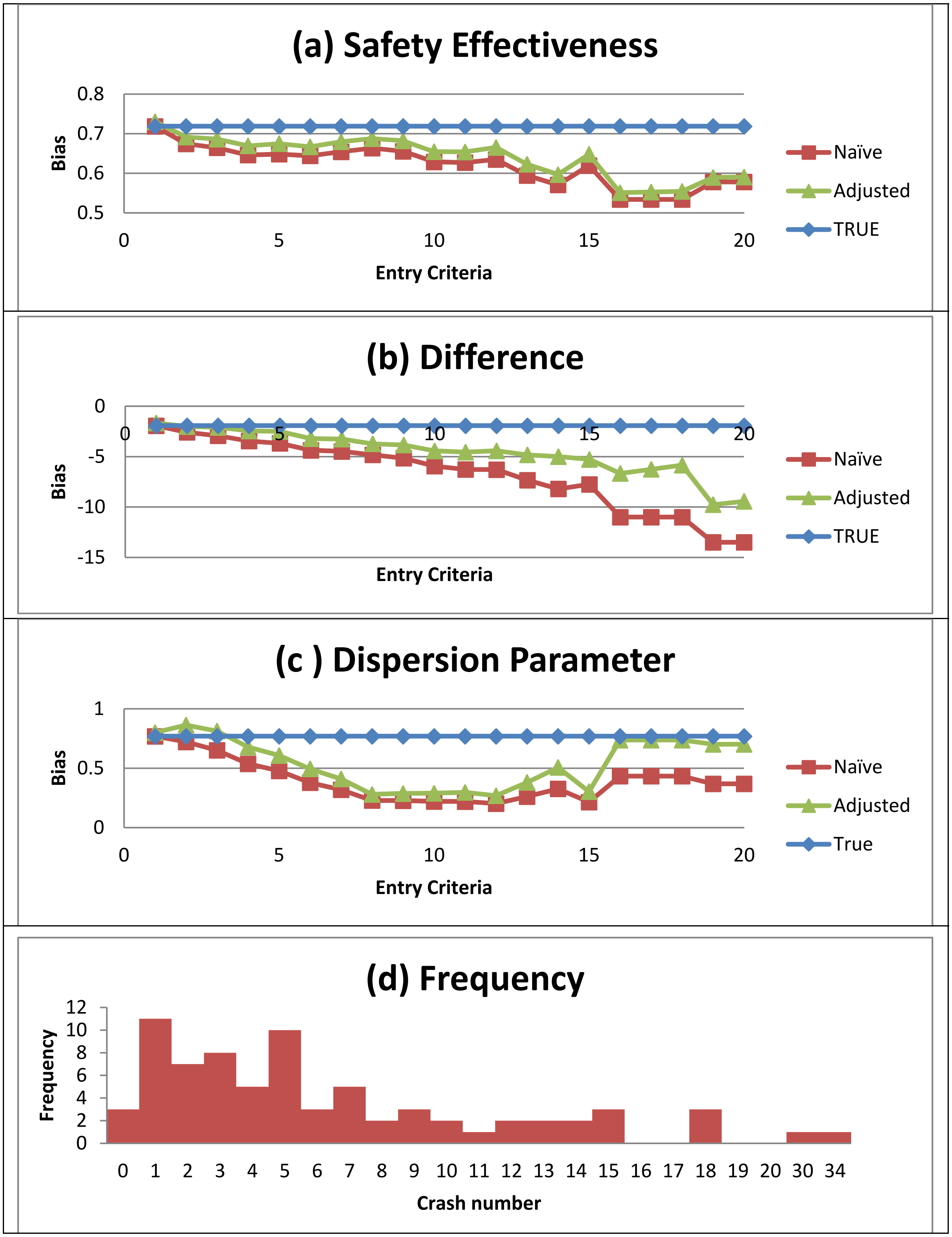

Fig. 5 Safety effectiveness, difference, and dispersion parameter for the true value, Naïve, and Adjusted methods

SUMMARY AND CONCLUSIONS

Previous studies have shown that the NAM provides a more precise estimate for the effectiveness of a treatment compared to the Naïve method, and is also better than the CG and the EB methods when similar control group data are not available (Lord and Kuo,2012; Kuo and Lord, 2013). Unfortunately, the above results were based on the assumption that all sites that are part of the treatment group have the same crash mean. This assumption may not be necessarily true when real or observed crash data are used. In this paper, we examined the accuracy of the NAM using datasets in which each site in the treatment group has a different mean value.

To evaluate how each of the biases would work in practice, crash data with a varying mean were simulated. Based on the simulated scenarios, it was shown that the NAM reduced the site-selection bias by approximately 50%, even when the crash mean values were characterized by large heterogeneity. However, it was pointed out that the estimator might overestimate the site-selection bias when the entry criteria are low. The three site-section bias estimators were also applied to two observed crash count datasets. The results supported the previous findings based on simulation data: setting higher entry criteria results in higher site-selection bias. Furthermore, estimators based on the NAM produced values closer to the true value than did the Naïve estimators for the safety index, the difference in mean values, and the dispersion parameter.

Similar to any other statistical methods, there are some limitations with the NAM method. Firstly, when the variance of the crash mean is large, then the benefits associated with the NAM method are reduced because the overall site selection bias is relatively small. Secondly, when the entry criteria are low or the sample size is relative small, the adjusted estimator may still be biased.

The Adjusted method can be applied to a wide range of before-after studies which have an entry criterion (or several criteria). Future studies should focus on applying the NAM for estimating site-selection biases for various types of data having different mean and sample-size values, such as traffic violations, driving conflicts, or speed data. As noted earlier, because of the hotspot identification guidelines and warrants for treatments from manuals, such as the MUTCD (FHWA, 2009), it is not surprising to note that most crash datasets from current before-after safety studies are truncated using an entry criterion, even if it is not explicitly defined. Future work should also focus on the accuracy/effectiveness of the Adjusted method when the real population mean, dispersion parameter, and site-selection bias are fully unknown (e.g., such as those used and published in previously available studies).

APPENDIX: NAM CALCULATION SPREADSHEET

An Excel spreadsheet to calculate NAM estimators is available, and interested readers may request it by email. In the spreadsheet below, input data are italicized, and output data are bold. The $\theta_{naive}, \Lambda_{1,naive}, \alpha_{naive}, \theta_{adjusted}$ were calculated based on the Equation (1) and (9) in Table 1. Note that the numbers in this calculation spreadsheet below were obtained by programs written in Excel, so the values are slightly different from our estimators documented in the main text that were produced from R (R Core Team, 2013).

Step 1. Enter the before and after crash data in the spreadsheet. By default, the spreadsheet will calculate the naïve estimator automatically.
Step 2. Calculate the parameters and probability and (1-cumulative probability) of the negative binomial distribution.
Step 3. Calculate the safety index based on the NAM (Equation 4). NAM values are shown in Appendix Table 1.

Site ID	Before Crash Data	After Crash Data	(after-mean after)^2
1	*5*	*5* Step 1	9.759
2	*5*	*4*	17.007
3	*5*	*0*	65.999
.	.	.	.
119	*29*	*20*	141.040
120	*31*	*41*	1080.833
121	*32*	*32*	570.065
sum	*1217*	*983*	
θ_{naive}	0.808	E((after-mean after)^2)	60.803
$\Lambda_{1,naive}$	10.058	8.124	
α_{naive}	0.798	prob_est	0.111
P(N>C+1)	0.616	Step 2	
P(N=C+1)	0.060		
$\theta_{adjusted}$	0.833	Step 3	

REFERENCES

1. Abbess, C., Jarrett, D., and Wright, C. (1981). "Accidents at black-spots: estimating the effectiveness of remedial treatment, with special reference to the regression-to-mean effect." *Traffic Engineering and Control*, 22, 535–542.
2. American Association of State Highway and Transportation Officials (AASHTO). (2010). *Highway Safety Manual,* Washington, D.C.
3. Austin, P. C. (2011). "An introduction to propensity score methods for reducing the effects of confounding in observational studies." *Multivariate behavioral research*, 46(3), 399-424.
4. Biondi-Zoccai, G., Romagnoli, E., Agostoni, P., et al. (2011) "Are propensity scores really superior to standard multivariable analysis?" *Contemporary Clinical Trials* 32 (5), 731-740.
5. Danielsson, S. (1986). "A comparison of two methods for estimating the effect of a countermeasure in the presence of regression effects." *Accident Analysis and Prevention* 18, 13–23.
6. Davis, G., (2000). "Accident reduction factors and causal inference in traffic safety studies: a review." *Accident Analysis and Prevention,* 32, 95–109.
7. FHWA, (2009). *Manual on Uniform Traffic Control Devices. Federal Highway Administration*, Washington, D.C. (Accessed Jul. 4th, 2013:http://mutcd.fhwa.dot.gov/pdfs/2009/pdf_index.htm)
8. Hamed, M.M., AlEideh, B.M., and AlSharif, M.M. (1999). "Traffic accidents under the effect of the Gulf crisis." *Safety Science,* 33, 59–68.
9. Hauer, E. (1980a). "Bias-by-selection: overestimation of the effectiveness of safety countermeasures caused by the process of selection for treatment." *Accident Analysis and Prevention*, 12, 113–117.
10. Hauer, E. (1980b). "Selection for treatment as a source of bias in before-and-after studies." *Traffic Engineering and Control,* 21, 419–421.
11. Hauer, E. (1997). *Observational Before-After Studies in Road Safety*. Pergamon Publications, London.
12. Hauer, E., Byer, P., and Joksch, H. (1983). "Bias-by-selection: the Accuracy of an unbiased estimator." *Accident Analysis and Prevention*, 15, 323–328.
13. Johnson, N.L., Kotz, S., and Balakrishnan, N. (1970). *Distributions in statistics: continuous univariate distributions-1*. Houghton Mifflin, Boston, MA
14. Kuo, P.-F., and Lord, D. (2013). "Accounting for Site-Selection Bias in Before-After Studies for Continuous Distributions: Characteristics and Application Using Speed Data." *Transp. Res. Part A: Policy Pract.,* 49, 256-269.

15. Lee, S., Choi, J., & Kim, S. W. (2010). Bayesian approach with the power prior for road safety analysis. Transportmetrica, 6(1), 39-51.
16. Lim, S., Marcus, S. M., Singh, T. P., Harris, T. G., & Seligson, A. L. (2014). Bias due to sample selection in propensity score matching for a supportive housing program evaluation in New York City. PloS one, 9(10), e109112.
17. Lord, D., and Kuo P-F. (2012). “Examining the Effects of Site Selection Criteria for Evaluating the Effectiveness of Traffic Safety Improvement Countermeasures.” *Accident Analysis & Prevention*, 47, 52-63.
18. Lord, D. (2006). “Modeling Motor Vehicle Crashes using Poisson-gamma Models: Examining the Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter.” *Accident Analysis & Prevention,* 38(4), 751-766.
19. Lord, D., S.P. Washington, and Ivan, J.N. (2005). “Poisson, Poisson-Gamma and Zero Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory.” *Accident Analysis & Prevention*. 37(1), 35-46.
20. Maher, M., Mountain, L. (2009). “The sensitivity of estimates of regression to the mean.” *Accident Analysis and Prevention,* 4, 861–868.
21. Miranda Moreno, L., Fu, L., Ukkusuri, S., and Lord, D. (2009). “How to incorporate accident severity and vehicle occupancy into the hotspot identification process?” *Transportation Research Record,* 2102, 53–60.
22. Park, E. S., Park, J., & Lomax, T. J. (2010). A fully Bayesian multivariate approach to before–after safety evaluation. Accident Analysis & Prevention, 42(4), 1118-1127.
23. Persaud, B. N. (2001). Statistical methods in highway safety analysis. NCHRP Synthesis 295. Transportation Research Board, Washington, D.C.
24. Persaud, B. N. and Lyon, C. (2007). “Empirical Bayes before-after safety studies: lessons learned from two decades of experience and future directions.” *Accident Analysis and Prevention*, 39, 546-55.
25. R Development Core Team, (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.
26. Rock, S.M. (1995). ”Impact of the 65 mph speed limit on accidents, deaths, and Injuries in Illinois.” *Accident Analysis and Prevention*, 27, 207–214.
27. Sacchi, E., & Sayed, T. (2015). Investigating the accuracy of Bayesian techniques for before–after safety studies: The case of a “no treatment” evaluation. Accident Analysis & Prevention, 78, 138-145.

28. Shin, K., & Washington, S. P. (2012). Empirical Bayes method in the study of traffic safety via heterogeneous negative multinomial model. Transportmetrica, 8(2), 131-147.
29. Tse, L. Y., Hung, W. T., & Sumalee, A. (2014). Bus lane safety implications: a case study in Hong Kong. Transportmetrica A: Transport Science, 10(2), 140-159.
30. Winkelmayer, W. C., & Kurth, T. (2004). Propensity scores: help or hype?. Nephrology Dialysis Transplantation, 19(7), 1671-1673.
31. Wood, J. S., Gooch, J. P., & Donnell, E. T. (2015). Estimating the safety effects of lane widths on urban streets in Nebraska using the propensity scores-potential outcomes framework. Accident Analysis & Prevention, 82, 180-191.
32. Wood, J. S., & Donnell, E. T. (2017). Causal inference framework for generalizable safety effect estimates. Accident Analysis & Prevention, 104, 74-87.
33. Wright, C., Abbess, C., and Jarrett, D. (1988). "Estimating the regression-to-mean effect associated with road accident black spot treatment: towards a more realistic approach." *Accident Analysis and Prevention*, 20, 199–214.