Investigating the Development of CMFs from Probability Analyses

Final Report

Report: ATLAS-2018-24

by

Raul Avelar, PhD, PE
Associate Research Engineer
Texas A&M Transportation Institute

Dominique Lord, PhD
Associate Transportation Researcher
Texas A&M Transportation Institute

and

Sruthi Ashraf
Texas A&M Transportation Institute

University of Michigan
2901 Baxter Rd
Ann Arbor, MI 48109-2150

Texas A&M University
Texas A&M Transportation Institute
College Station, TX 77843-3135

August 2018

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

ACKNOWLEDGMENTS

This work was sponsored by the Advancing Transportation Leadership and Safety Program (ATLAS) at the Texas A&M Transportation Institute (TTI). The ATLAS Center is supported by a grant from the U.S. Department of Transportation, Office of the Assistant Secretary for Research and Transportation, University Transportation Centers Program (DTRT13-G-UTC54). The ATLAS Center is a collaboration between the University of Michigan Transportation Research Institute and the Texas A&M Transportation Institute.

The authors would like to thank Tomas Lindheimer for his assistance with the literature review organization and editing; and Bahar Dadashova for her assistance in the preliminary stages of developing the synthetic data sets for analysis. The authors would also like to thank Robert Wunderlich and Barb Lorenz for their support throughout this project.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
ATLAS-2018-24		

4. Title and Subtitle	5. Report Date
Investigation on the Development of CMFs from Probability Analyses	August 2018
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Raul Avelar, Dominique Lord, and Sruthi Ashraf	

9. Performing Organization Name and Address	10. Work Unit no. (TRAIS)
Texas A&M Transportation Institute The Texas A&M University System College Station, TX 77843-3135	11. Contract or Grant No. DTRT13-G-UTC54

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Advancing Transportation Leadership and Safety (ATLAS) Center 2901 Baxter Rd., Room 124, Ann Arbor, MI 48109-2150 USA	14. Sponsoring Agency Code

15. Supplementary Notes

Supported by a grant from the U.S. Department of Transportation, OST-R, University Transportation Centers Program

16. Abstract

The *Highway Safety Manual* (HSM) defines a Crash Modification Factor (CMF) as the ratio of the number of crashes expected after a modification or measure implementation to the number of crashes per unit of time estimated if the change does not take place. The success of applying methods such as the cross-sectional, and before-after analyses to estimate CMFs is highly dependent on the size of the data available and specific characteristics about the crashes under study. Significant challenges are present when evaluating the safety performance of improvements in the face of limited crash data or extremely rare crashes. A representative sample of locations is needed for a robust crash frequency safety analysis, in general, drawing a representative sample of sites will result in many sampled sites (if not all sites) having zero crashes under low crash frequency conditions. Fortunately, probability-based analysis is applicable in those cases. However, it is not completely clear how a safety effect estimated from these types of analyses relates to the definition of a CMF in the HSM that is based on crash frequency, not crash risk. This research investigated the feasibility of developing reliable CMFs for countermeasures using risk analysis.

Through the development of a theoretical framework that links the odds ratio (OR) derived from probability-based analysis (PBA) and the true CMF, researchers proposed an alternative estimator of the CMF derived from PBA (CMF_PBA) consisting of a modification of the OR based on the base odds of crashes in the population of sites under study. Researchers demonstrated that this modified estimator is not dependent on the sampling procedure. Under the theoretical framework, researchers showed that the OR should generally exhibit bias from the true CMF. One would draw over-optimistic results if the true CMF were smaller than one or over-pessimistic results if the CMF were larger than one. Researchers used simulated scenarios to test the performance of the two probability-based estimates (CMF_PBA and OR) and a comparable frequency-based analysis estimate (CMF_FBA). The results of these analysis showed that the CMF_FBA is in general the best estimate of the three, except in the case of rare crashes (i.e., the expectation of a crash is very small). In that case, researchers found the CMF_FBA to exhibit a significant positive bias when estimating small values of the true CMF, and that both the OR and the CMF_PBA were robust against that bias. Researchers recommend the use retrospective analysis in those cases, and the adoption of the CMF_PBA over the OR when estimating roadway safety effects since the CMF_PBA performed significantly better over the OR across an ample set of scenarios, as found from the analysis of the simulation results.

17. Key Words	18. Distribution Statement
Crash Modification Factor, CMF, Retrospective Analysis, odds ratio	Unlimited

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	64	

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS, ACRONYMS, AND INITIALISMS

AADT	Annual Average Daily Traffic
CI	Confidence Interval
CMF	Crash Modification Factor
CMF_FBA	CMF Estimator derived from frequency-based analysis
CMF_PBA	CMF Estimator derived from probability-based analysis
CRA	Complete Records Analysis
FBA	Frequency-Based Analysis
FHWA	Federal Highway Administration
HSM	Highway Safety Manual
i.i.d.	Independent and Identically Distributed
MLE	Maximum Likelihood Estimation
OR	Odds Ratio
PBA	Probability-Based Analysis
PMLE	Penalized Maximum Likelihood
SE	Standard Error
SEIF	SE Inflation Factor

CHAPTER 1. INTRODUCTION

The power attainable by statistical analyses is bound by sample size, the size of the safety effect, probability distribution, correlations between variables of analysis, and the range represented by key variables. Probability analysis is a viable alternative to analyze safety in situations with limited crash data. For example, wrong-way crashes are extremely rare, a situation that makes it difficult to directly measure the potential crash reduction of countermeasures. In order to perform a safety evaluation in these cases, it is desirable to use every site with known history of crashes occurring and supplement these sites with a set of control sites for comparison. However, by evaluating sites that were chosen based on their history of crashes, biased results are likely to ensue. Even when supplementing the study with sites with no history of crashes, bias is only reduced in the best-case scenario when traditional analysis techniques are used.

Prospective versus Retrospective Analyses

Data sets collected based on the response variable are known as retrospective or case-control data sets. As mentioned earlier, these data sets are useful to handle situations where the available number of cases or controls is limited. In contrast, a prospective analysis is preferred when sufficient data are available, where sites are selected based on their representativeness of the variables of interest (e.g., presence/absence of countermeasures) and the values of the response variable (number of crashes) are obtained after site selection for analysis. Irrespective of the nature of the data set (retrospective or prospective), the intent of a probability analysis is to estimate how the probability of a crash changes as a function of covariates of interest. In this case, the most common analytical tool is some variant of logistic regression, which estimates the odd ratios for risk factors associated with the probability of a crash. The actual proportion of crashes on a sample of data highly depends on how the data were collected (either prospectively or retrospectively). Figure 1 illustrates the typical differences in distributions of crashes that result from sampling prospectively (on the left) and retrospectively (on the right).

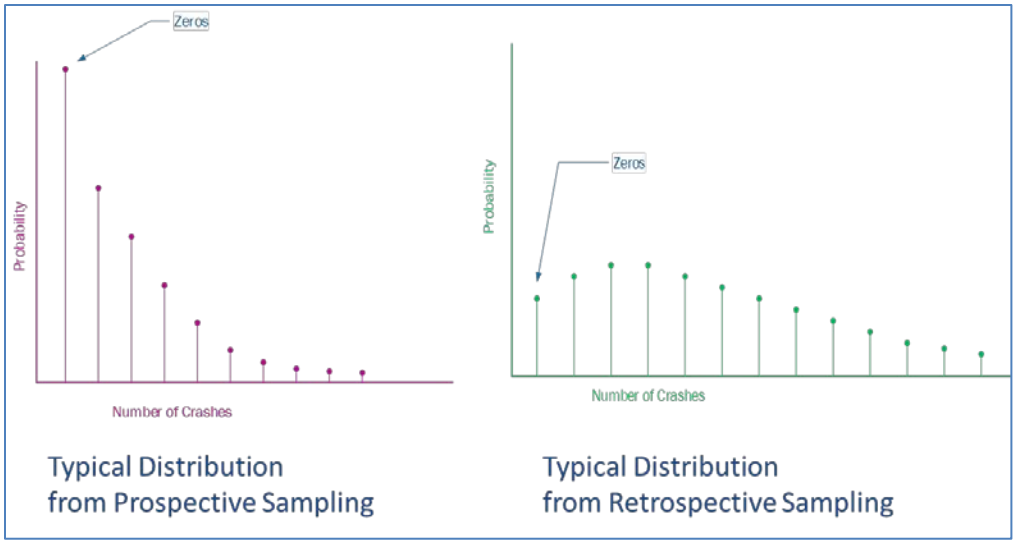


Figure 1. Prospective and Retrospective Distributions from a Population of Sites in Common.

This distortion of the true distribution of crashes in retrospective sampling have the potential to severely bias the results from analyses that focus on changes of the distribution parameters, such as Poisson and Negative Binomial crash-frequency models. However, the effect of an explanatory variable on the response variable is not sensitive to the sampling scheme or the distortion in distribution shape if such effect can be expressed and interpreted as a change in the odds of the response variable, rather than a change in mean expectation. The impact of different sampling schemes should only be evident on the intercept term of the logistic regression results [1].

The results from applying the logistic regression method, despite being robust to the sampling scheme, in general do not directly translate to the scale of expected crashes. A measure of such expectation would require that the distribution of crashes in the complete sample (including control cases) be considered representative of any facility similar to those in the study. While frequency analysis results can be easily used to formulate crash modification factors (CMFs) compatible with the framework of the *Highway Safety Manual* (HSM) [2], the literature only indicates empirical or anecdotal evidence of comparable (possibly interchangeable) CMFs developed from probability-based analyses. For example, previous work [3] has shown that the CMF estimated from a case-control setup follows closely the trends of known CMFs from the HSM. Other approaches include previous works [4] that combined results from probability, exposure, and frequency analysis to estimate the expected number of truck collisions with bridge piers. Avelar et al. [5] developed probability models to estimate the risk of collisions with tire debris left on roadway pavement surface due to the unfeasibility of developing crash frequency models for such a rare type of crash.

Despite these and other prior efforts, the relationship between results from risk and frequency analyses is not completely understood. For instance, in the section for case-control studies of the reference document *A Guide to Developing Quality Crash Modification Factors* [6], it is stated, "the odds ratio is a direct estimate of the CMF." In the same section of the report, it is stated that "case-control studies cannot be used to measure the probability of an event (e.g., crash, severe injury) in terms of expected frequency." Given the current knowledge about CMFs developed from retrospective samples, this project assessed the degree of representativeness of these CMFs and the degree to which they can be deemed reliable based on a close examination of the mathematical statistics of both prospective and retrospective analyses.

Project Objectives

The primary goal of this study was to develop a methodology, if feasible, that obtained reliable CMFs and their standard errors (SEs) from probability-based safety evaluations that are equivalent to the CMFs expected from a frequency-based analysis. The specific objectives were as follows:

1. To establish the relationship of equivalence between the outcomes of frequency-based analyses (FBA) and probability-based analyses (PBA) for safety evaluations.
2. To develop a methodology to construct CMFs and SEs from PBA safety evaluations.
3. To test and validate the methodology using a synthetic data set.

Structure of the Report

This report is divided into five chapters. Chapter 1 introduces this research and the structure of the report. Chapter 2 summarizes the literature review performed prior to designing an analysis plan and evaluation. Chapter 3 develops a theoretical framework that links the results from retrospective analysis to the equivalent effect on the crash frequency domain. Chapter 4 summarizes the data development process, experiment design, and analysis. Finally, Chapter 5 provides the conclusions and outlines the future directions from this work.

CHAPTER 2. LITERATURE REVIEW

This chapter summarizes literature relevant to this research's questions of interest. Researchers performed a thorough literature search on CMF development within the last 20 years. The first section presents an overview of CMFs in the context of the HSM. The next two sections describe different methods that have been proposed for estimating CMFs. The second section of this chapter describes the basic methods that been used for their development. The third section covers recent methods that have been proposed for handling small and incomplete data sets for the estimation of CMFs. The description also provides information about the strengths and weaknesses of the proposed methods. A final subsection is added to discuss the findings from the literature.

Introduction

CMFs are statistical estimates that describe how key variables, such as geometric and operational variables, affect the occurrence of crashes on the facilities. In some safety-related documents, CMFs are defined as a measure of the estimated effectiveness of a safety countermeasure [7]. However, more broadly, a CMF can be defined as a change in crashes due to a shift in one or more variables that can either positive (reduction) or negative (increase). A CMF is used as a multiplicative factor to compute the expected number of crashes at a location after implementing a specific countermeasure or a change in geometric or operational variables. A CMF with a value of more than one indicates an expected increase in the number of crashes, whereas a CMF with a value less than one indicates a decrease in the number of crashes. CMFs can be presented either as a single value (point estimate) or a function that considers relevant site characteristics [7].

The HSM and the Highway Safety Improvement Program manual provide guidelines to identify and prioritize sites, select appropriate countermeasures for safety enhancement of those sites, and determine the effectiveness of these safety countermeasures. The HSM provides a catalog of CMFs for various geometric and operational treatment types, which are based on robust scientific evidence [8]. The CMF Clearinghouse is a web-based database, managed by the Federal Highway Administration (FHWA), which provides a very comprehensive list of CMFs that safety analysts and practitioners can use for different types of projects. As CMFs are developed under the assumption that all other conditions and site characteristics remain constant, the validity of CMFs rely on consistent and agreeable base conditions.

Traditional Development of Crash Modification Factors

FHWA provided a guide in 2010 for developing CMFs [9]. Table 1 summarizes all study designs in this document.

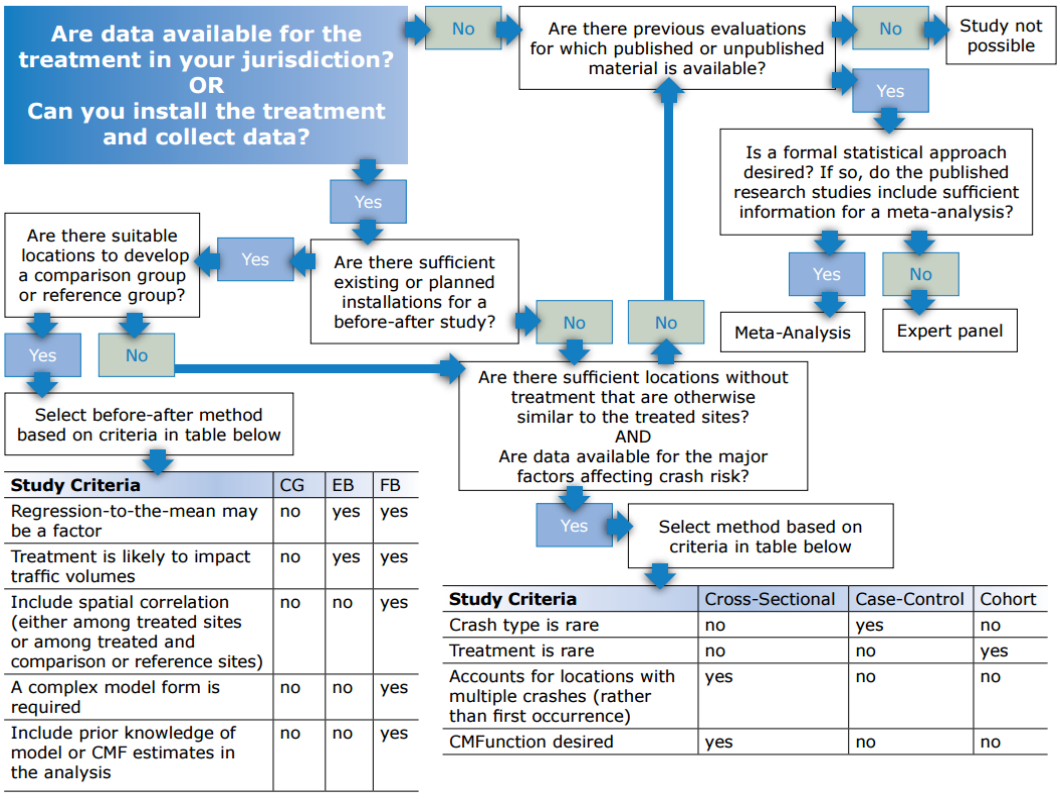
Table 1. Summary of Study Designs for Developing CMFs [9].

Study Design	General Applicability	Strengths	Weaknesses
Before-After w/Comparison Group	Treatment is similar among treatment sites. Untreated sites are used to account for non-treatment related crash trends.	Accounts for non-treatment related time trends and changes in traffic volume.	Difficulty in accounting for regression-to-the-mean.
Before-After w/Empirical Bayes	Treatment is similar among treatment sites. A separate comparison group may be required where the treatment affects the reference group.	Accounts for regression-to-the-mean, traffic volume changes, non-treatment related time trends.	Cannot include prior knowledge of treatment, considerations for spatial correlation, or complex model forms.
Full Bayes	Used in before-after or cross-section studies. Useful when complex model forms are required or samples are small. Previous CMFs are to be introduced in the modeling.	Reliable results with small sample sizes. Can include prior knowledge, spatial correlation, and complex model forms in the evaluation.	Implementation requires a high degree of training.
Cross-Sectional	Requires sufficient sites that are similar except for the treatment of interest.	Useful when predicting crashes. Allows estimation of CMFs when conversions are rare.	CMFs may be inaccurate if applied without care. Omitted variable bias, correlation among variables, and inappropriate functional form are some of the potential issues.
Case-Control	Assess whether exposure to a potential treatment is disproportionately distributed between sites with and without the target crash.	Useful for studying rare events because the number of cases and controls is predetermined.	Can only investigate one outcome per sample. Does not differentiate between locations with one crash or multiple crashes.
Cohort	Used to estimate relative risk, which indicates the expected percent change in the probability of an outcome given a unit change in the treatment.	Useful for studying rare treatments because the sample is selected based on treatment status.	Only analyzes the time to the first crash. Large samples are often required.
Meta-Analysis	Combines knowledge on CMFs from multiple previous studies while accounting for the study quality in a systematic and quantitative way.	Can be used to develop CMFs when data are not available for recent installations and it is not feasible to install the strategy and collect data. Can combine knowledge from several jurisdictions and studies.	Requires the identification of previous studies for a particular strategy. Requires a formal statistical process. All studies included should be similar in terms of data used, outcome measure, and study methodology.
Expert Panel	Expert panels are assembled to evaluate the findings of published and unpublished research critically. A CMF recommendation is made based on agreement among panel members.	Can be used to develop CMFs when data are not available. Can combine knowledge from several jurisdictions and studies.	Traditional expert panels do not systematically derive precision estimates of a CMF. Possible forecasting bias.
Surrogate Measures	Surrogate measures may be used to derive a CMF where crash data are not available or insufficient.	Can be used to develop CMFs in the absence of crash-based data.	The approach to establish relationships between surrogates and crashes is relatively undeveloped.

Some of the highlights are as follows:

1. For before-after studies, data need to be available for both treated and untreated sites.
2. A cross-sectional study is useful when before-after data are limited.
3. Case-control studies indicate the likelihood of crash reduction/increase by a treatment by using odds ratio (OR) analysis.
4. For expert panels, a formal statistical process is not required.

The selection of a study design depends on whether there are data on the treatment available or whether the treatment can be installed, and the data can be collected after the installation. The document presents a flow chart to help engineers select the appropriate study design for developing one or more CMFs (Figure 2).



Study Criteria	CG	EB	FB
Regression-to-the-mean may be a factor	no	yes	yes
Treatment is likely to impact traffic volumes	no	yes	yes
Include spatial correlation (either among treated sites or among treated and comparison or reference sites)	no	no	yes
A complex model form is required	no	no	yes
Include prior knowledge of model or CMF estimates in the analysis	no	no	yes

Study Criteria	Cross-Sectional	Case-Control	Cohort
Crash type is rare	no	yes	no
Treatment is rare	no	no	yes
Accounts for locations with multiple crashes (rather than first occurrence)	yes	no	no
CMFunction desired	yes	no	no

Figure 2. Flow Chart for Study Design Selection [9].

If there are no data available or collecting data after installation of the treatment is unfeasible, then a meta-analysis or an expert panel may be used to develop the CMF. If there are sufficient data, but no sufficient similar locations without the treatment and data are available for the major factors affecting crash risk, then cross-sectional, case-control, and cohort study methods may be used. Gross and Donnell suggested that case-control and cross-sectional studies produce consistent results if care is taken in the study design and the development of

regression models (i.e., having many covariates, functional form). Case-control and cross-sectional studies may provide a viable alternative to estimate CMFs when a before-after study is impractical due to data restrictions. For example, when CMFs for fixed roadway lighting (at-grade intersections, Minnesota) and the allocation of lane and shoulder widths (rural two-lane highway segments, Pennsylvania) were estimated, the case-control method produced CMFs that were greater than respective CMFs from the cross-sectional method. By comparing the results from full and restricted models, the study illustrated the importance of controlling for potential confounding variables while estimating safety effects and the danger of excluding important covariates, which may result in over- or under-estimation of the expected safety effects [10]. The quality of a CMF is related not only to the study design but also to other factors including sample size, robustness of data, SE, and accounting for potential sources of bias. Table 2 shows the sample size considerations for each study design.

For cross-sectional studies, more data are required when variables of interest are not statistically significant. A traditional before-after analysis can be performed with data collected from 10 to 20 sites. In comparison, cross-sectional studies require data including at least 100, but often 1,000 or more sites or observations. The standard deviation is a better indicator of the precision of the CMF when safety effectiveness of the treatment for a future application in a different area or jurisdiction is to be determined [11]. Increasing the sample size will not necessarily reduce the standard deviation but improving (or reducing the uncertainty for) the relationship between the CMF and the risk factors associated with the crash modification function. Quality data for crashes and traffic volume are crucial in developing CMFs.

Advanced Methods for Developing CMFs

Other methods have been proposed for developing CMFs in response to the need for large samples, or large data collection efforts required by traditional methods. For example, one method combined the safety effects of various CMFs, since CMFs may not be independent in practice [7]. Park and Abdel-Aty combined multiple CMFs using a weighed linear regression model and an analytic hierarchy process by considering different roadway types and crash severity levels [12]. The authors concluded that the devised method could overcome over-estimation of CMFs and produce results that are more reliable when safety effects of multiple treatments are combined.

Table 2. Sample Size Considerations for Methods Adopted to Develop CMFs.

Method	Sample Size Considerations
Before-After Studies with Comparison Group	The variables that impact the precision (SE) with which the CMF is estimated are: 1. The size of the treatment group, in terms of the number of crashes in the before period. 2. The relative duration of the before and after periods. 3. The likely (postulated) CMF value. 4. The size of the comparison group in terms of the number of crashes in the before and after periods.
Before-After Studies with Empirical Bayes	Currently, there is no formal method for determining required sample sizes for the empirical Bayes before-after approach. The method presented by Hauer can be used to approximate the sample size required. The sample size estimates could be considered conservative as the empirical Bayes approach reduces uncertainty in the estimate of expected crashes [13].
Full Bayes	Sample size considerations for full Bayes modeling are similar to those for cross-sectional studies or before-after studies.
Cross-Sectional	For cross-sectional studies, the number of locations required will depend on several factors including: 1. Average crash frequencies. 2. The number of variables desired in the model. 3. The level of statistical significance desired in the model. 4. The amount of variation in each variable of interest between locations.
Case-Control	The required sample size for a case-control study design is calculated by using an equation that considers the case to control ratio, prevalence of treatment, statistical significance level, and the common proportion over two groups among other factors.
Cohort Studies	For a cohort design, the required sample size may be calculated using the ratio of treatment group to reference group, proportion in the reference group where an outcome was observed, desired detectable relative risk (i.e., magnitude of the safety effect to be detected), and common proportion over two groups among other factors.
Meta-Analysis	NA (weighting of CMFs) [14]
Expert Panel	NA (expert panel is similar to the meta-analysis approach but is less formal)
Surrogate Measures	NA (based on availability of reliable model)

A method that has been used to avoid the need of large-scale data collection for cross-sectional studies was the application of safety surrogate measures by Saleem and Lorion [15]. The study used traffic conflicts as a safety surrogate to evaluate the effectiveness of several safety treatments. The treatments were: installing left-turn lanes on major road approaches at four-leg signalized intersections, installing left-turn lanes on major road approaches at three-leg two-way stop controlled intersections, installing right-turn lanes on major road approaches at four-leg signalized intersections, and changing the left-turn signal control from permissive to protected-permissive at four-leg signalized intersections. Using traffic simulation software, researchers observed the changes in traffic conflicts due to changes in road design and computed Conflict Modification Factors. Crash-conflict relationship was then used to estimate CMFs. The CMFs were then compared to the CMFs available from the HSM and the CMF Clearinghouse. The proposed methodology yields CMFs that closely match CMFs obtained from observational studies [15].

Probability analysis using logistic regression techniques, OR, and other univariate statistical analyses have been applied to study the probability increase in crash risk involving the use of alcohol and the use of a cellular phone [16, 17]. Researchers in Australia examined the correlation between variables and sleep-related crashes on slow-speed roads. Sleep-related vehicle crashes made up 1.6 percent of total crash data on low-speed roads. The data were analyzed using Chi-square tests with Cramer's V as an estimate of the effect along with a multivariate analysis consisting of a series of logistic regressions [18].

A crash occurrence is typically considered a rare event, and especially rare in some cases. Researchers have postulated that working with rare events in general tends to exacerbate the bias [19]. In logistic regression, maximum likelihood estimates (MLE) are consistent but only asymptotically unbiased [20]. Exact logistic regression foregoes asymptotic properties of estimates as with the MLE. Because the total event count matters, not percentage, exact logistic regression may be used when sample size is too small compared to using the usual logistic regression estimated using the MLE [21]. Exact logistic regression is useful when sample size is less than 200, the covariates are discrete, and the total number of covariates is small. Firth has proposed a bias correction method, known as the penalized maximum likelihood estimates (PMLE) [22]. The PMLE were found to be unbiased, even in cases with small sample size, as the estimates always converged and solved the problem of separation [23]. Researchers recommended to keep the sample size large and apply PMLE when estimating logistic regression models with rare events data [24]. Some data scientists have paired modeling rare events based on logistic regression model with post-hoc adjustments of the results to cast predictions applicable to the population level [25]. For example, the post-hoc adjustment proposed by Tanish is based on a simple Ordinary Least Squares regression developed from the logistic regression predictions and a few deciles of those proportions obtained directly from the marginal distribution of the response in the population.

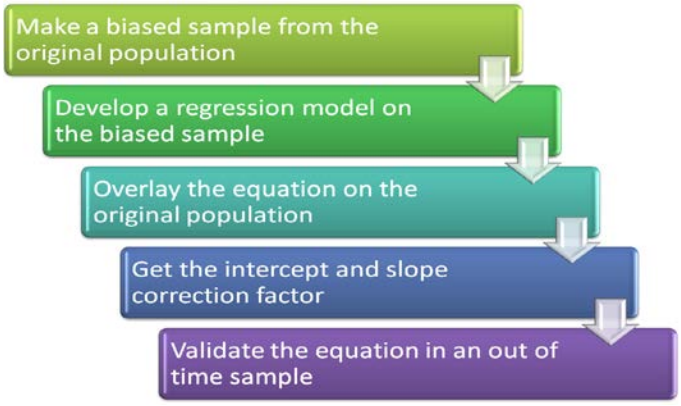


Figure 3. Step by Step Process of Modeling Rare Events Based on Usual Logistic Regression Model [25].

To estimate the likelihood of future crashes, Das et al. employed the logistic regression method using eight years of traffic crash data in Louisiana. The authors conducted an exploratory analysis to develop a crash prediction model that would estimate the likelihood of future crashes for at-fault drivers, and in particular, those defined as crash-prone drivers. Researchers analyzed over 700,000 crash records that had a driver declared to be at-fault. Drivers involved in multiple crashes in a single year and marked as at-fault were classified as crash-prone. Researchers estimated that 5 percent of licensed Louisiana drivers are crash-prone. The model considered 371 attributes from the crash records, including driver age, roadway lighting condition, and weather. According to the authors, logistic regression is particularly beneficial when analyzing a data set that contains many explanatory variables [26].

Researchers in Greece considered road crashes as rare-events when analyzing traffic data at the Attica Tollway. Researchers proposed a series of rare-events logit models to analyze road accident occurrences and used real-time traffic data from three random loop detectors in the Attica tollway located in the Greater Athens area. The study concludes by stating that the logit model provides an adequate statistical fit to estimate the relationship between crash occurrence and traffic factors, such as speed [27].

King and Zeng preferred choice based or endogenous stratified sampling (case-control) data collection strategies over random sampling or exogenous stratified sampling (cohort/cross-sectional) strategies when crashes were considered as rare events [28]. King and Zeng used rare event logistic regression to address problems in the statistical analysis of rare events data (under 5 percent of total event data). The regression analysis provided an unbiased approach based on absolute and relative risks. Maximization of the weighted log-likelihood was adopted instead of traditional maximizing log-likelihood. Parameters were estimated based on weighted least-squares regression, which were said to reduce the bias and variance. Researchers recommend case-cohort study when the study includes a variable that is expensive to collect. For all ranges of subsampling, the approximate Bayesian estimate produced a lower root mean square error than the logit and unbiased estimator. The correction method proposed by King and Zeng somewhat over-corrects the bias in MLEs as sample size gets small (200 or smaller) [29].

Veazey et al. [30] used the rare event logistic regression proposed by King and Zeng [28] to predict the distribution of mesophotic hard corals across the main Hawaiian Islands and implemented it using the Zelig package in R with satisfactory results. Researchers documented their workflow well, showing many precautions to avoid well-known modeling pitfalls. A correlation scatter plot matrix was constructed, and highly correlated variables were excluded. Predictors that lacked a clear distribution pattern or correlated minimally were also excluded. The inclusion of squared terms in the regression equation permitted the logistic curve to reflect

a bell curve shape expected from a preliminary exploration of the data. All possible combinations of included covariates were modeled and ranked using the corrected Akaike information criterion. Researchers checked small-scale, local spatial autocorrelation using Geary's C statistic, based on the deviations in the responses of observation points with one another. The authors also computed the global spatial autocorrelation using the Moran's I statistic, which measures cross-products of deviations from the mean value. Receiver operating characteristics curves were plotted, and overall prediction success of each model was calculated. Values for sensitivity and specificity for threshold increments of 0.005±1 standard deviation of the rounded mean for each model were calculated.

The ORs for a given exposure are routinely obtained within logistic models while controlling for confounders. They are often interpreted as equivalent to relative risks, which may lead to potentially serious problems, as researchers in public health argue that the OR always overestimates relative risk [31]. Roadway safety researchers have argued that case-control studies cannot be used to measure the probability of an event (e.g., crash, severe injury) relative to the expected frequency and are more often used to show the relative effects of risk factors [32]. These researchers advocate for multiple logistic regression techniques can be used to clarify these relationships because of the logistic regression's ability to examine the risk associated with one factor while controlling for other factors [32].

Shrestha et al. used multiple logistic regression with a limited amount of data in Nevada. Researchers investigated the factors associated with vehicle crash severities in built-up areas along rural highways in Nevada. The study used a binary logistic regression model to analyze a total of 337 crashes that took place around 11 towns from 2002 to 2010, and an OR analysis to calculate the risk likelihood related to the severity of the crash. Researchers also calculated the marginal effects of the covariates. The study found reliable predictors using this method. Speeding and inattention were associated with a greater likelihood of more severe crash [33].

Klauer et al. conducted an in-depth analyses of driver inattention using data collected from 100 vehicles in a naturalistic driving study. For the analysis, the study created an event database (i.e., crashes, near-crashes, incidents) and a baseline database. The baseline database was created by stratifying the large amount of data according to number of crashes, near-crashes, and incidents in which each vehicle was involved. After the stratification, a sample of 20,000 baseline epochs made up the baseline database. The baseline database was used as a case-control data set for more accurate calculation of OR. The study found that handheld device use, driver drowsiness, and engaging in complex secondary tasks increase the crash/near-crash risk [34]. Assumption bias may overstate the results risk obtained from an OR analysis. Young reanalyzed the data used in a study by Virginia Tech Transportation Institute. Substantially lower estimates of population exposure percent (P_e %) and population attributable risk percent

were obtained when using a standard method for epidemiological analysis when compared to the method adopted by Virginia Tech Transportation Institute. The author suggested that the bias might be due in part to the definition of variables and other data reduction techniques used before performing the OR analysis [35].

Abdel-Aty et al. modeled the crash probability using a logistic model by constructing conditional likelihood and matched case-control logistic regression based on real-time traffic flow data obtained from loop detectors. The estimated parameters were log OR, which can be used to approximate the relative risk of a crash and develop a prediction model under the matched crash/non-crash analysis by establishing a threshold value that yields desirable crash classification accuracy. To identify all significant variables, a binary outcome variable was modeled using the stratified conditional logistic regression method. Higher occupancy rates upstream coupled with high variation in speed downstream of the crash location, both at 5 to 10 min before the crash; increase the likelihood of crash occurrence at a location in-between [36].

Crash risk analysis models were developed under the assumption that the same traffic status would hold identical crash probability for different roadway sections. A study by Yu, Wang, and Abdel-Aty adopted a hybrid latent class analysis modeling approach to address heterogeneous effects of geometric characteristics. The proposed model was tested using Shanghai's urban expressway crash data, geometric characteristics data, and roadway section traffic data [37]. Separate crash risk analysis models were established for the crashes at the three homogeneous subgroups (based on geometric characteristics) using Bayesian random parameter logistic regression models once the crash data were segmented to three classes based on crash occurrence locations depending on latent class analysis results. The optimal number of latent classes were found using bootstrap likelihood ratio tests. The parameters were estimated by maximum likelihood using expectation-maximization procedure [38].

Wu et al. developed a new CMF for rural two-lane undivided horizontal curve data using a cross-sectional study design. The study compared the CMF findings to those in the HSM and the CMFs developed by Gooch et al. [39]. The curves were divided into eight bins based on the frequency of curves and each bin contained approximately similar number of curves. Mean radii of curves within a bin was considered as the primary feature and the total number of observed crashes represented the safety of curves in that bin. The number of crashes for each curve was predicted by considering them tangents using the base HSM safety performance function for rural two-lane highways. Initial ratios of observed to predicted safety was calculated and normalized based on the bin with greatest radii. The results showed that decreasing radii increases the number of crashes and the developed methodology outperformed others [40].

In 2016, Gooch et al. used a mixed-effects negative binomial regression to quantify the safety performance of horizontal curves on two-way, two-lane rural roads. To overcome the limitations of small sample sizes and unreported SEs, researchers used a propensity score-potential outcomes framework. The propensity scores-potential methodology was used to estimate a CMF for horizontal curves. A cross-sectional analysis was used to explore the impacts of adjacent curves on crash frequency. The model used eight years of crash data obtained from 10,000 miles of two-lane rural roads in Pennsylvania. The study found that presence of horizontal curves was associated with an increase in crash frequency. The CMF for total crash frequency was a function of the degree of curvature and curve radius. Overall, the CMF increased as curve radius decreased. The study also found a decrease in crash frequencies when adjacent curves were close [39].

Crash records may also be incomplete or have missing data fields. A Complete Records Analysis (CRA) logistic regression was used in a study to estimate exposure ORs without bias for studies with missing data. The CRA was found to be asymptotically unbiased, if the data are missing at random in large samples, meaning the missing record is independent of the variables involved in the analysis. Exposure ORs were estimated without bias (asymptotically) when missing data were not jointly dependent on exposure and outcome. The authors recommended the use of directed acyclic graph to check whether the missing data mechanism falls within the classes where CRA logistic-regression exposure OR estimate are asymptotically unbiased [41]. Other studies recommend a sensitivity analyses when a large proportion of data is missing [42].

Summary

This review of literature documented different methods for estimating CMFs. The traditional methods usually require large data sets, the use of a control or reference group, and complete set of crash records. Lacking the complete data sets may lead to an overestimation of risk factor and an inaccurate CMF for a particular treatment. On the other hand, a small sample size may also lead to a biased estimate of the CMF. In order to overcome these limitations, researchers have used different statistical methods to account for small sample size and missing data. The literature review described those methods that have been proposed in the past.

Specific to logistic regression approaches, this review found various nuances and issues that researchers, both in the roadway safety and other arenas, have identified when retrospective samples are used to estimate risk. The next chapter delves into the specific theoretical issues that emerge when logistic regression is applied to a hypothetical crash generating process.

CHAPTER 3. THEORETICAL FRAMEWORK

This chapter develops a theoretical framework to establish analytically how a safety estimate derived from a retrospective analysis relates to the CMF as defined in the HSM. The framework is developed starting from the crash generation process and accepted premises applied to the hypothetical that the CMF definition in the HSM is true.

Crash Generation Process

This section outlines principles believed to underlie the process by which crashes occur (named the crash generation process from this point on).

Homogeneous Poisson Process

A homogeneous Poisson process can be defined in several ways. Most relevant to highway safety, it is a count process of events (i.e., crashes) emerging from a succession of independent and identically distributed (i.i.d.) Bernoulli trials from a crash-generating process with parameter p. A second, oftentimes useful definition is that of a stationary and independent process whose events are independent and the time periods between events. As a result, the count of events obtained from a fixed time period is a Poisson distributed random variable. An important property of this homogeneous process is that it is memoryless (independent events in the dimension of time). Another important property is that the convolution of independent Poisson processes is a Poisson process itself.

Non-Homogeneous Poisson Process

A non-homogeneous Poisson process is a Poisson process for which the single distribution parameter of the count process is a variable itself. A potentially useful way to handle non-homogeneous Poisson processes is by defining them as a homogeneous Poisson process on a non-linear time scale, which permits the application of some useful known properties of homogeneous Poisson processes in handling the non-homogeneous case. In this context, imposing a probability distribution to describe the variability of the count process parameter in the data generation process could be seen as a scaling of the time between the events.

Time between Events in Poisson Processes

In a homogeneous Poisson process, the times between events are independent, so it can be shown that the time between events are i.i.d. exponential random variables. In the case of a heterogeneous Poisson process, characterizing the stochasticity of the time between events is not straightforward. However, in this case, a scale transformation exists such that the times between events are Geometric i.i.d. random variables in the transformed time scale as mentioned before.

Emergence of the Poisson Distribution

Because a homogeneous Poisson process generates independent times between events, the process is essentially a set of successive independent Bernoulli trials in a given period of time Δt. Each trial takes place in an epoch δ resulting from a homogeneous partition of Δt into n equal sub periods. In this case, the expectation of the number of events during Δt is the sum of the expectations during all subperiods:

$$E(N) = \lambda = n \cdot p \qquad \textbf{Equation 1}$$

Where,

$E(N)$ = Expected value of N (number of crashes);

λ = Mean parameter;

n = Number of trials in experiment; and

p = Probability of crash per trial.

Where *N* is the number of events in Δt, λ is the expectation for *N*, *n* is the number of equal subperiods in Δt, and *p* is the expectation of success in each Bernoulli trial. An implied assumption here is that not more than one event is likely to occur in one epoch, if at all (a reasonable assumption for a large enough *n* with respect to λ). The probability of obtaining *y*, a given value of *N,* can be written using the Binomial distribution:

$$P(N = y) = \binom{n}{y} \cdot p^y \cdot (1-p)^{(n-y)} \qquad \textbf{Equation 2}$$

Where y is a realization of random variable N. It is straight forward to demonstrate that, for a fixed success rate p (i.e., from a homogeneous Poisson process), the following relation is exact:

$$P(N = y) = \lim_{n\to\infty}\left[\binom{n}{y} \cdot p^y \cdot (1-p)^{(n-y)}\right] = \frac{\lambda^y}{y!} \cdot e^{-\lambda} \qquad \textbf{Equation 3}$$

which corresponds to the Poisson distribution.

Discussion

Previous work has shown that deviations from the assumption of a constant Bernoulli expectation in the homogeneous Poisson process explains overdispersion to a great extent (i.e., convex function of the mean of the count) typically observed in crash data [43]. A random change in the expectation of each Bernoulli trial would in turn result in random changes in the count process expectation, which would result in a non-homogeneous Poisson process (i.e.,

probability a crash varies from trial to trial in the crash generating process). Such a process lends itself to modeling as some Poisson mixture model.

Modeling the Crash Generating Process: Exposure and Direct Proportionality

As indicated in Equation 1, the expectation of the count function is directly proportional to a number of trials n and some probability determined by other factors. The number of trials can be understood as the exposure to the Poisson process. Strictly speaking, exposure refers to a length in time that the Poisson process is at work, so that a Poisson distribution can emerge from a repeated binomial experiment, as shown in the previous section. However, exposure can refer to dimensions other than time. For example, the length of a segment must have a direct relationship to the crashes within that segment.

AADT as Exposure

In the two prior examples of exposure (i.e., time and length), the direct proportionality does not entail any curvature (i.e., the constant exponent for the exposure variable in a model must be 1.0). In early years, such type of proportionality was expected also between crashes and the annual average daily traffic (AADT) variable, but the results of several works have shown that not to be the case in general [2, 44, 45, 46, 47, 48, 49]. There are many potential explanations of this contrast: study setup, endogeneity in the model, heterogeneous units of analysis, and differences in operational characteristics that may affect safety, among others. Regardless of other potentially influential factors, the AADT is a yearly average and as such, it is not really exposure to traffic but a reasonable surrogate. In other words, because two sites equal in their geometries and even with the same AADT can have very different hourly traffic patterns, the AADT is not a direct measure of exposure for any given small epoch. Therefore, the anticipated exponent of AADT as exposure has been found at times statistically significantly different from 1.0, either larger or smaller than one by different research studies.

Crash Generating Process from a Spatial-Temporal Exposure Standpoint

Considering exposure in time and in space, as discussed above, let the rate of a non-homogeneous Poisson process be defined locally in small differentials of space and time so that the expectation is approximately fixed, for a given level of traffic exposure expressed by a surrogate measure, such as the AADT. In that case, the crash expectation at this limited window of time and space is defined as:

$$E(N) = \lambda = \Delta t \cdot \Delta S \cdot f(AADT) \cdot \lambda_0 \qquad \textbf{Equation 4}$$

Where,

Δt	=	Small amount of time exposure.
ΔS	=	Small amount of space exposure.

$f(AADT)$	=	Non-linear monotonic functional form of traffic exposure using AADT as a surrogate.

All other variables as previously defined.

Sampling Scheme and CMF Estimation

Note the definition of the CMF according to the HSM and its relation to the sampling scheme (either prospective or retrospective). Equation 5 represents the definition of a CMF in terms of crash expectation, per the HSM:

$$CMF = \frac{E(N|C=1)}{E(N|C=0)} \qquad \textbf{Equation 5}$$

where,

CMF	=	Crash Modification Function/Factor.
N	=	Number of Crashes.
C	=	Indicator variable for a Condition, 1 if present, 0 otherwise.

For a given period and location, and after controlling for other relevant exposure, the CMF represents the multiplicative change in crash expectation when a treatment is present (i.e., C=1) relative to the treatment not being present (i.e., C=0). Therefore:

$$CMF = \frac{\lambda_{C=1}}{\lambda_{C=0}} = \frac{p_{C=1}}{p_{C=0}} \qquad \textbf{Equation 6}$$

The adequacy of Equation 6 depends on the reasonable assumption that true exposure (either time or space, per Equation 4) should cancel when defining this ratio. If this is not a reasonable assumption, then 1) the CMF value must be dependent on the amount of exposure; and 2) the application of the CMF should affect the relationship between crashes and exposure.

Re-expressing the above definition in terms of probabilities conditional to the treatment:

$$CMF = \frac{p_{C=1}}{p_{C=0}} = \frac{P(N>0|C=1)}{P(N>0|C=0)} \qquad \textbf{Equation 7}$$

Equation 7 will be revisited after the next section that introduces and discusses retrospective sampling.

Retrospective Sampling

A retrospective sampling scheme entails a sample that is selected based on the output variable. This approach is advantageous in cases of especially rare events where there is interest in using all known events in the analysis. In this case, an analyst would supplement the data set of events known with an additional data set representing non-events. Logistic regression can handle these situations, as the OR has been shown to be invariant between retrospective and prospective samples [1]. In this context, p in Equation 1 refers to the marginal prospective probability of an event, which relates directly to the unconditional probability of the count variable (as shown in Equation 3). In the case of a retrospective sample, however, the count variable is conditional to the sampling scheme and therefore:

$$P(N = y|Z = 1) \neq P(N = y) \qquad \textbf{Equation 8}$$

Where,

Z	=	Indicator variable equals 1 when the unit is included in sample, 0 otherwise.
$E(Variable\|Condition)$	=	The conditional expectation of Variable given a Condition.

Other variables as previously defined.

A relationship exists between the prospective and retrospective probabilities. Let the sampling rate κ of a retrospective sampling scheme be defined as the ratio shown in Equation 9:

$$\kappa = \frac{P(Z = 1|N = 0)}{P(Z = 1|N > 0)} \qquad \textbf{Equation 9}$$

κ represents the odds of non-cases to be included in the sample and it is calculated as the ratio of the number of non-cases to the number of cases included in a given sample. For a sampling scheme with known κ, the prospective and retrospective probabilities relate as shown in Equation 10:

$$p = \frac{\kappa \cdot p^*}{1 + \kappa \cdot p^* - p^*} \qquad \textbf{Equation 10}$$

Where,

p	=	Prospective probability that N>0;
p^*	=	Retrospective probability that N>0;

Other variables as previously defined.

The definition in Equation 9 relates the sampling procedure (via the outcome distribution conditional to the sample) and the marginal distribution of the outcome variable in the population under study (i.e., the unconditional outcome distribution). However, after applying the Bayes theorem, κ can be re-expressed in terms of the conditional distribution as observed in the sample and the marginal distribution of the crashes in the population as shown next in Equation 11:

$$\kappa = \frac{\frac{P(N=0|Z=1)}{P(N>0|Z=1)}}{\frac{P(N=0)}{P(N>0)}} \qquad \textbf{Equation 11}$$

After making the corresponding substitutions, it is straightforward to show how starting from Equation 11 one can arrive at Equation 10, or alternatively to Equation 12.

$$p^* = \frac{\frac{1}{\kappa} \cdot p}{1 + \frac{1}{\kappa} p - p} \qquad \textbf{Equation 12}$$

Relationship between the CMF and the OR

Considering all the above relationships, and under the assumption that the HSM definition of CMF is correct, Theorem 3.1 shown next encapsulates the relationship between the frequency-based CMF and the OR obtained from the analysis of a retrospective sample.

Theorem 3.1. The Relationship between the CMF and the OR, Given a Retrospective Sampling Scheme

For a given retrospective analysis with sampling rate κ that empirically estimates $OR_{C=1}$, the relationship of this OR estimate and the true CMF is independent of the sampling scheme and governed by Equation 13:

$$CMF = OR_{C=1} \cdot \frac{1 + Odds_{C=0}}{1 + Odds_{C=0} \cdot OR_{C=1}} \qquad \textbf{Equation 13}$$

where,

$Odds_{C=0}$	=	The unconditional odds of a crash when the treatment is absent.
$OR_{C=1}$	=	The odds ratio for a crash when the condition is present, with respect to the condition not present.

Other variables as previously defined.

Proof of Theorem 3.1

Substituting Equation 10 in Equation 6:

$$CMF = \frac{p_{C=1}}{p_{C=0}} = \frac{\frac{\kappa \cdot p^*_{C=1}}{1 + \kappa \cdot p^*_{C=1} - p^*_{C=1}}}{\frac{\kappa \cdot p^*_{C=0}}{1 + \kappa \cdot p^*_{C=0} - p^*_{C=0}}}$$

After changing the representation of $p^*_{C=1}$ and $p^*_{C=0}$ in terms of odds, a complicated expression emerges:

$$CMF = \frac{\frac{Odds^*_{C=1}}{1 + Odds^*_{C=1}}}{1 + \kappa \cdot \frac{Odds^*_{C=1}}{1 + Odds^*_{C=1}} - \frac{Odds^*_{C=1}}{1 + Odds^*_{C=1}}} \Big/ \frac{\frac{Odds^*_{C=0}}{1 + Odds^*_{C=0}}}{1 + \kappa \cdot \frac{Odds^*_{C=0}}{1 + Odds^*_{C=0}} - \frac{Odds^*_{C=0}}{1 + Odds^*_{C=0}}}$$

After simplifying, one arrives at the following form:

$$CMF = \frac{\frac{Odds^*_{C=1}}{1 + Odds^*_{C=1}}}{\frac{Odds^*_{C=0}}{1 + Odds^*_{C=0}}} \cdot \frac{1 + \kappa \cdot \frac{Odds^*_{C=0}}{1 + Odds^*_{C=0}} - \frac{Odds^*_{C=0}}{1 + Odds^*_{C=0}}}{1 + \kappa \cdot \frac{Odds^*_{C=1}}{1 + Odds^*_{C=1}} - \frac{Odds^*_{C=1}}{1 + Odds^*_{C=1}}}$$

After the following substitution $Odds^*_{C=1} = Odds^*_{C=0} \cdot OR_{C=1}$ one arrives at the following expression after further simplification:

$$CMF = OR_{C=1} \cdot \frac{[1 + Odds^*_{C=0} + Odds^*_{C=0} \cdot (\kappa - 1)]}{[1 + Odds^*_{C=0} \cdot OR_{C=1} + Odds^*_{C=0} \cdot OR_{C=1} \cdot (\kappa - 1)]}$$

Additional simplification yields the following expression:

$$CMF = OR_{C=1} \cdot \frac{1 + \kappa \cdot Odds^*_{C=0}}{1 + OR_{C=1} \cdot \kappa \cdot Odds^*_{C=0}} \qquad \textbf{Equation 14}$$

Equation 14 includes the retrospective base odds and the sampling rate κ and thus is dependent of the sampling scheme. Nonetheless, this equation may be useful when estimating the CMF from a retrospective analysis and the sampling rate is known. Although κ and $Odds^*_{C=0}$ are both dependent of the characteristics of the sampling scheme, they are related and inversely proportional. Examining Equation 12, it is evident that as $\kappa \to 0$, $p^* = P(N > 0|Z = 1) \to 1$ and so $Odds^*_{C=0} \to \infty$.

Starting from Equation 11, one can establish at this point the relationship between κ and $Odds^{*}_{C=0}$ conditional to each level of the indicator variable C. Because the sampling is carried independently of the variable C, the relation in Equation 11 remains unchanged when expressed conditional to either of the levels of the variable C (by the definition of statistical independence). Therefore, the two following expressions (Equation 15 and Equation 16) are true for sampling strictly based on the response variable. When re-expressing Equation 11 in terms of odds at C=0:

$$\kappa = \frac{\frac{P(N=0|Z=1)}{P(N>0|Z=1)}}{\frac{P(N=0)}{P(N>0)}} = \frac{\frac{1}{Odds_{C=0,\ Z=1}}}{\frac{1}{Odds_{C=0}}}$$

$$\kappa = \frac{Odds_{C=0}}{Odds_{C=0,\ Z=1}} \qquad \textbf{Equation 15}$$

As indicated above, Equation 11 can also be written conditional to C=1 and will remain unchanged due to independence of the sampling from variable C:

$$\kappa = \frac{Odds_{C=1}}{Odds_{C=1,\ Z=1}} \qquad \textbf{Equation 16}$$

After replacing Equation 15 and Equation 16 in Equation 14:

$$CMF = OR_{T=1} \cdot \frac{1 + \frac{Odds_{C=0}}{Odds_{C=0,Z=1}} \cdot Odds_{C=0,Z=1}}{1 + \frac{Odds_{C=0}}{Odds_{C=0,Z=1}} \cdot Odds_{C=0,Z=1} \cdot OR_{T=1}}$$

One arrives to Equation 17, which is the same as in Equation 13 (what was being proven):

$$CMF = OR_{C=1} \cdot \frac{1 + Odds_{C=0}}{1 + Odds_{C=0} \cdot OR_{C=1}} \qquad \textbf{Equation 17}$$

The relationship in Theorem 3.1 (i.e., Equation 13 or Equation 17) establishes that the odds at the population base condition, the OR, and the frequency-based CMF are related independently of the sampling scheme. Furthermore, as $Odds_{C=0} \to 0$, $OR_{C=1} \to CMF$ in Equation 17, which is the long time assumption in the transportation community about the OR and CMF equivalence.

Bias and Other Limitations of the OR when Estimating a Frequency-Based CMF

Beginning from Equation 17, the OR can be expressed as a function of the CMF and the base odds as shown in Equation 18:

$$OR_{C=1} = \frac{CMF}{1 + Odds_{C=0} - CMF \cdot Odds_{C=0}}$$ **Equation 18**

It is evident that if CMF>1, OR is also larger than 1, because in that case the denominator is smaller than 1:

$$1 + Odds_{C=0} - CMF \cdot Odds_{C=0} < 1$$

$$1 + Odds_{C=0} < 1 + CMF \cdot Odds_{C=0}$$

$$Odds_{C=0} < CMF \cdot Odds_{C=0}$$

$$CMF > 1$$

By similar reasoning, one can demonstrate that if CMF<1, then also OR<1. However, two corollaries result from Equation 18:

- If the CMF is larger than one, the OR is biased to the right (i.e., larger than the CMF).
- The OR is biased to the left for CMFs smaller than one.

The magnitude of the bias is directly proportional to the size of the base odds in the analysis. These two corollaries are demonstrated next.

Starting from the assumption that $OR_{C=1} > 1$, one can begin from the following assumption about the relative values of OR and CMF:

$$OR_{C=1} \overset{?}{>} CMF$$

$$\frac{CMF}{1 + Odds_{C=0} - CMF \cdot Odds_{C=0}} \overset{?}{>} CMF$$

$$\frac{1}{1 + Odds_{C=0} - CMF \cdot Odds_{C=0}} \overset{?}{>} 1$$

$$1 \overset{?}{>} 1 + Odds_{C=0} - CMF \cdot Odds_{C=0}$$

$$CMF \cdot Odds_{C=0} + 1 \overset{?}{>} 1 + Odds_{C=0}$$

$$CMF \cdot Odds_{C=0} \overset{?}{>} Odds_{C=0}$$

$$CMF \overset{?}{>} 1$$

which is true since the original premise was that $OR_{C=1} > 1$. Therefore, the OR is always larger than the CMF when CMF>1. By similar reasoning, if CMF<1, then the OR is always smaller than the CMF.

Figure 4 shows a plot of the relationship in Equation 18 for CMF<1.0. For this example, the value of CMF was fixed at 0.9. The resulting concave curve was as expected.

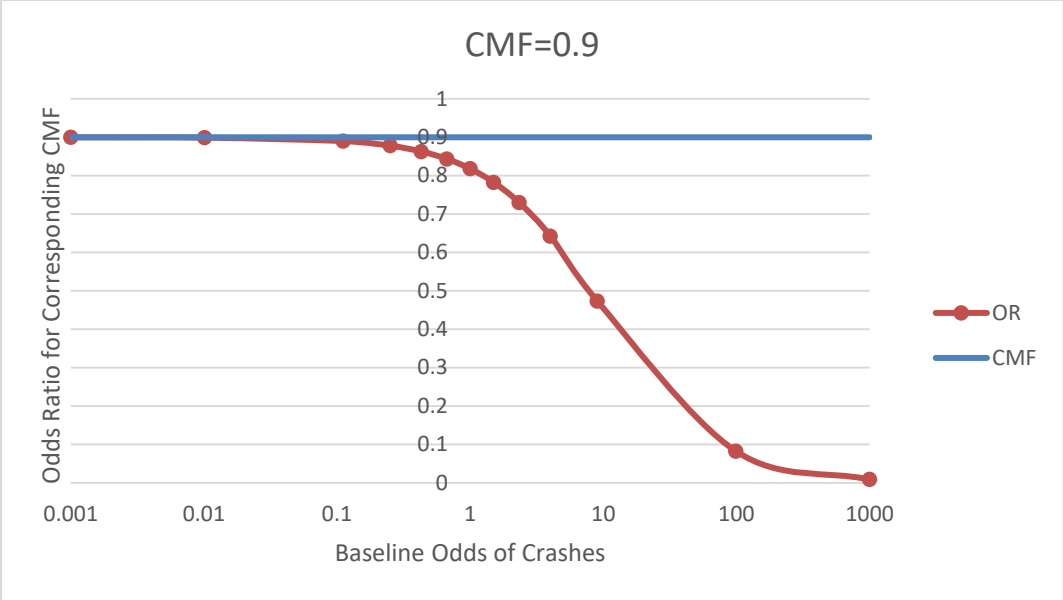


Figure 4. OR vs. Base Odds for CMF= 0.9.

As anticipated, the left bias of the OR increases (underestimating the CMF more severely) as the base odds increases. Similarly, Figure 5 shows the case of CMF>1 where the trend is that the OR degenerates into a convex line with increasing base odds.

An additional implication of Equation 18 is that it establishes a clear limitation of the OR to estimating the CMF as events become more and more common. Since the denominator in Equation 18 must be positive to have a physical interpretation, researchers can impose the following constraint:

$$1 + Odds_{C=0} - CMF \cdot Odds_{C=0} > 0$$

After rearranging the inequality, one arrives at Equation 19:

$$CMF < \frac{1}{Odds_{C=0}} + 1 \qquad \textbf{Equation 19}$$

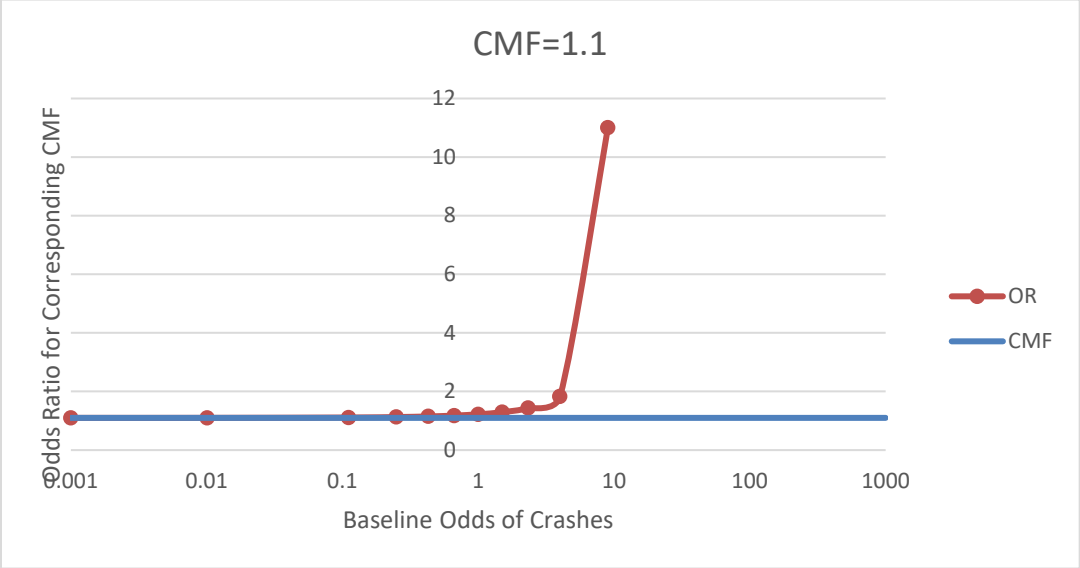


Figure 5. OR vs. Base Odds for CMF= 1.1.

Equation 19 establishes that the CMF is bound to be smaller than a positive, monotonically decaying function of the base odds. Table 3 shows the maximum estimable CMF for multiple values of base odds.

Table 3. Maximum Estimable CMF, Given Base Odds.

Odds (Crash |C=0)	P(crash | C=0)	Maximum Estimable CMF
0.05	0.05	21.00
0.1	0.09	11.00
0.25	0.20	5.00
0.5	0.33	3.00
1	0.50	2.00
2	0.67	1.50
4	0.80	1.25
8	0.89	1.13
16	0.94	1.06

Although Equation 19 imposes a potential limitation, such limitation could be circumvented by changing the base level for the analysis. In other words, if one intends to estimate a CMF larger than the limiting value per Equation 19, the analysis can be carried defining the base condition as the base (C=0), in which case the analysis would estimate CMF^{-1} without trouble.

Estimating the Standard Error of a CMF Derived from a PBA

This section derives the SE for a CMF estimated from Theorem 3.1 (Equation 13). The right-hand side of that equation has two estimates that are to be derived from a logistic regression analysis. Therefore, the SE estimate of interest is the SE of a function of random variables. For ease of calculation, it is convenient to break down Equation 13 in two subfunctions. Let the subfunctions A and B be defined as follows:

$$A = OR_{C=1} + OR_{C=1} \cdot Odds_{C=0}$$

$$B = 1 + Odds_{C=0} \cdot OR_{C=1}$$

Therefore, from multivariate calculus one can write:

$$\Delta CMF = \Delta A \frac{1}{B} - \Delta B \frac{A}{B^2}$$

Let β_0 and $\beta_{c=1}$ be the logistic regression estimates for the base odds and the OR, respectively, then:

$$\Delta A = \Delta\beta_{c=1} * \exp\beta_{c=1} + \Delta(\beta_{c=1} + \beta_0) * \exp(\beta_{c=1} + \beta_0)$$

$$\Delta B = \Delta(\beta_{c=1} + \beta_0) * \exp(\beta_{c=1} + \beta_0)$$

The expression $\Delta\beta_{c=1}$ is obtained directly from the information matrix of the logistic model fit and the estimate of $\Delta(\beta_{c=1} + \beta_0)$ can be obtained from the simple operation:

$$SE(\beta_{c=1} + \beta_0) = \sqrt{SE(\beta_0)^2 + SE(\beta_{c=1})^2 + 2 \cdot Cov(\beta_{c=1} + \beta_0)}$$

After the appropriate substitutions, one arrives to the expression in Equation 20 for the CMF SE:

$$\begin{aligned} &SE[CMF] \\ &= \left\{\frac{\exp\beta_{c=1}\left(SE[\beta_{c=1}] + \exp(\beta_0) \cdot \sqrt{SE(\beta_0)^2 + SE(\beta_{c=1})^2 + 2 \cdot Cov(\beta_{c=1} + \beta_0)}\right)}{1 + \exp(\beta_{c=1} + \beta_0)}\right\} \\ &- \left\{\frac{(\exp(2\beta_{c=1} + \beta_0) + \exp(2\beta_{c=1} + 2\beta_0)) . \sqrt{SE(\beta_0)^2 + SE(\beta_{c=1})^2 + 2 \cdot Cov(\beta_{c=1} + \beta_0)}}{(1 + \exp(\beta_{c=1} + \beta_0))^2}\right\} \end{aligned} \quad \textbf{Equation 20}$$

Estimating the Standard Error of a CMF Derived from a PBA

In the case of additional covariates, the meaning of the base odds becomes unclear (as it is conditional to the base level of the additional covariates) and was explored in the next chapter.

In terms of SEs for those cases, researchers estimated $SE[Odds_{C=0}]$ as SEs for functions of random variables themselves, similar to the derivations in the last section.

Chapter Summary

This chapter explored the relationship between the CMF and the OR from a theoretical standpoint. No distributional assumptions were made to keep the results of the exploration as general as possible. The important implications of this analytical evaluation are listed as follows:

- There is an upper bound of the estimable CMF via PBA. Such upper bound decreases as the base odds increase toward infinity.
- For CMFs larger than one, the OR is always larger than the CMF, which implies a right bias. Such right bias is a function of the base odds. As the base odds increase toward infinity, so does the OR.
- Conversely, for CMFs smaller than one, the OR is always smaller than the CMF, which implies a left bias. Such left bias is again a function of the base odds. As the base odds increase toward infinity, the OR converges toward zero.
- Despite the found biases of the OR with respect to the CMF, the OR converges toward the CMF as the base odds decrease toward zero.

CHAPTER 4. ANALYSIS

This chapter summarizes a battery of tests performed on synthetic data on the comparative performance of CMF estimation via FBA and PBA. Researchers performed calculations of estimates and their precision using the appropriate expressions derived in the prior chapter.

Synthetic Data Set Development

Researchers wrote computer code to implement the statistical specification of the two-way-two-lane undivided rural-highway crash prediction model as defined in the HSM. The intent to implement this code was to simulate the crash generation process that model describes, so that each simulation yielded the number of crashes per year for a set of conditions fixed experimentally. The code requires specifying influential roadway characteristics in a synthetic location that it then passes on to the crash generation process, which returns a random realization of the number of crashes in a year—a positive random value between zero and infinity— that follows the probability distribution established by the HSM two-way, two-lane highway in the HSM.

The computer code described above allowed researchers to generate multiple synthetic (but realistic) data sets to explore the performance of CMF estimation under various scenarios. The size of the data set to be generated, for example, was an input variable that researchers controlled to observe the varying performance of PBAs at different sample sizes.

In addition to AADT and a set of fixed standard geometric design features (i.e., 12 ft lanes, 6 ft paved shoulders, and 0.5 miles of length), each synthetic data set included a subset (about 30 percent of sites) with a simulated safety intervention, whose CMF was also controlled by researchers. This way, researchers could observe the performance of the analyses for different values of CMFs.

Matching Sampling Scheme

As demonstrated in Chapter 3, the bias of the OR in estimating a CMF depends on how rare the events under study are. For all practical purposes, the OR and the CMF are equivalent for the rarest events, and they tend to diverge as the events become more and more common. In turn, the rarity of crashes is in direct inverse relation to the amount of exposure in the crash-generating process. In most studies, it is practical to include locations with a range of AADTs, which quantify exposure to traffic. To represent this feature from real studies, researchers generated synthetic data sets that included a range of AADT values, so the data sets included a mix of crash propensities that varied with AADT. This feature, although made for a more realistic set of scenarios, required some way to account for crash risk varying systematically across the range of AADT. Typically, one includes a feasible functional form of the variable of

interest among the regressors as a covariate of the variable of interest in regression analyses. The inclusion of additional variables, in turn, is deemed to affect the intercept term significantly, as this term is adjusted to the mean reference level of all regressors. Since the intercept term captures the base odds, it is unclear what the implications are when shifting such term as a function of model adjustments due to other variables. Researchers included various alternative adjustment to obtain base odds while accounting for this issue where appropriate.

To explore performance across various such base odds adjustments, researchers constructed a matched retrospective sample from each synthetic data set, such that the range of AADTs of the rarest instance in the data set (either sites with crashes or sites without crashes) was matched according to the sampling rate k. Researchers then explored the performance of various PBAs including a range of complexity dealing with the base odds issue. The next section briefly describes the adjustments proposed and tested in this research.

Proposed Adjustments for Base Odds

As mentioned earlier, some decisions should be made about the base odds because of the presence of other variables influential to safety in the analysis. Particularly, crash risk is well known to increase with increasing AADT, so AADT is an important confounder to be considered. How this variable it is treated in the model should cause a significant impact on the estimated base odds as discussed before. The following are the cases that were considered in this evaluation:

Case 1. **Base Odds and OR are the only estimands in the PBA, no adjustment to the sample base odds.** This is the simplest case and it relies in the matching of the data set to control implicitly for the confoundedness between the OR and the AADT. Because the AADT range and site distribution is exactly the same between the cases and the controls, the estimate of the OR should be orthogonal between the two subsets, and thus good estimators of the true OR and PBA CMF should be possible to construct. The estimator for the PBA CMF is constructed using Equation 13 with an assumption that the odds from the PBA are an acceptable estimate of the population odds. The PBA CMF from this case is anticipated to show bias as the sampling rate deviates from 1.0.

Case 2. **Base Odds and OR are the only estimands in the PBA and sample base odds are adjusted toward population base odds.** In this case, the analysis is the same as for case 1, but an adjustment of the sample base odds is performed via the property shown in Equation 14.

Case 3. **Base Odds and OR are the only estimands, the analysis is adjusted for AADT as an offset, and no adjustment is done to the sample base odds.** The only difference between this case and Case 1 is the inclusion of the natural logarithm of the AADT as an

offset to the PBA. The estimation of the CMF is carried on under the same assumption for the base odds in case 1 and should suffer of the same issues.

Case 4. **Base Odds and OR are the only estimands, the analysis is adjusted for AADT as an offset, and sample base odds are adjusted toward population base odds.** Similar to Case 2, only adjusting the sample base odds to reflect the population base odds from the sampling scheme and the data at hand.

Case 5. **Base Odds and OR are estimated alongside with an AADT trend, and no adjustment is done to the sample base odds.** This case and Case 3 both account for AADT, but with the difference that the magnitude of that relationship is added among the regression coefficients. However, the estimation of the CMF is carried under the same assumption for the base odds in Case 1 and Case 3.

Case 6. **Base Odds and OR are estimated alongside with an AADT trend, and an adjustment is done to the sample base odds toward the population base odds.** This case is the same as Case 5 except that an adjustment toward the population based odds is done by a weighted combination of the sample base odds estimate and the median AADT value in the sample. The weights are chosen such that the mean-value adjustment equals the sample-based estimate of the population base odds (estimation done via Equation 14). Although this procedure is rather complex, it is expected to yield a realistic SE for the adjusted CMF estimate because such SE incorporates the covariances between the OR, the sample base odds, and the AADT coefficient estimates.

Case 7. **Base Odds and OR are estimated alongside with an AADT trend, the sample base odds adjusted toward the population base odds with assumption of independence.** This case is very similar to Case 6 but with the adjustment toward the population base odds made through a marginal estimate rather than through a linear combination of the intercept and AADT coefficient estimates.

Case 8. **Comparable Prospective Sample.** This case consisted of a prospective analysis over an independent prospective sample of the same size of the other seven cases for comparison.

Given the definition of the eight cases, there are three models fitted to produce the cases, so the OR is the same for the pair Case 1 and Case 2; the pair Case 3 and Case 4 also share the same OR, as does the trio Case 5, Case 6, and Case 7. Researchers coded the three models as a categorical variable and an indicator variable for the population adjustment of the base odds. The next section describes the design of the experiment to test the performance of these outlined cases.

Experiment Design

In order to test the performance of the various cases, researchers considered six potentially important factors for the evaluations:

- Rareness of crashes. Three levels were chosen for this factor:
  - Rare events, average annual crash expectation per sample fixed at 0.05.
  - Typical events, average crash expectation per sample fixed at 3.
  - Common events, average crash expectation per sample fixed at 20.
- Randomness of crashes. Three levels were chosen for this factor:
  - Widely dispersed events, population inverse dispersion parameter Theta (of an NB distribution) in the population fixed at 2.
  - Common dispersion events, population dispersion parameter in the population fixed at 10.
  - Poisson-like events, population dispersion parameter in the population fixed at 100.
- Size of the CMF. Three levels were chosen for this factor:
  - Small, the population CMF was fixed at 0.25.
  - Middle, the population CMF was fixed at 0.75.
  - Large, the population CMF was fixed at 1.5.
- Sampling rate. Three levels tested here:
  - Heavy on controls, k=5.
  - Balanced, k=2.
  - Heavy on cases, k=0.2.
- Sample size. Three levels:
  - Small, number n=200 approximately.
  - Medium, n= 800 approximately.
  - Large, n=2500 approximately.
- Weighted regression. Two levels:
  - Weighted proportionally to the number of crashes.
  - No weights.

There are $3^5 \times 2 = 486$ combinations of these factors. Since there were two estimates per each of the eight cases described in the prior section—except for case 8, with only one estimate for the PBA CMF—researchers anticipated running a total of 7,290 estimates for analysis. However, the number of estimates was reduced because the sampling procedures were not always feasible due to issues that emerged from randomness in the synthetic data. For example, in some instances a small sample size, high rareness of crashes, and high randomness of crashes resulted in synthetic data sets that would not have enough control or treatment sites to satisfy the required sampling rate. This was particularly likely for the smallest and largest sampling rate values. In an effort to curb this issue, researchers first tried inverting the sampling rate in the

problematic cases (because the issue was essentially over-representing the rarer of the two types of sites). Researchers did not generate a synthetic data set in the cases that such inversion would still not yield a data set for analysis.

Due to the above issues, a total of 6,225 estimates were obtained from a total of 415 statistically independent runs (as opposed to the originally anticipated 486 runs).

Measures of Effectiveness

Researchers defined two measures of effectiveness to compare the relative performance of the seven retrospective cases and their performance relative to the prospective analysis defined for that purpose (i.e., Case 8). The following a brief description of the two measures of effectiveness for the analysis:

- Probability of capturing the parameter. This is a binary variable that takes the value of 1 if a 95 percent confidence interval (CI) around the estimate contains the CMF of interest, zero otherwise.
- SE Inflation with respect to Case 8. This is the ratio of the SE of each estimate to the SE of the independent FBA on a prospective synthetic data set of comparable size.

Results

After generating the 6,225 estimates as described in the prior section, researchers proceeded to an examination of the trends. Such exploratory analysis is shown in the next subsection.

Exploratory Analysis

To provide a sense of the results obtained from each run, Table 4 shows a sample result for a run where the set of 15 estimates resulted from one synthetic data sets obtained from a fixed set of factors per the table header.

From this table, only estimates from cases 1 through 4 are successful in capturing the true parameter (i.e., CMF=0.25) when constructing 95 percent CIs for the OR. Interestingly, the cases where the OR fails to capture the parameter are those that included the additional AADT regression estimate. In contrast, the true parameter is captured in all seven cases where the CMF is estimated from the combination of the OR and the base odds. However, the drawback to that success is that the SEs of the estimate tend to be larger for the PBA CMF estimates, in some cases three to four times as large as their OR counterparts.

Table 4. Sample Set of Runs for p=0.47, Dispersion=0.2, CMF=0.25, k=2, n=796, and No Weights.

Case	OR	OR.SE	CMF_PBA	CMF_PBA.SE
C1	0.1925	0.0447	0.2856	0.0607
C2	0.1925	0.0447	0.4331	0.1304
C3	0.1925	0.0447	0.2856	0.0607
C4	0.1925	0.0447	0.4331	0.1304
C5	0.1567	0.0377	0.1905	0.0490
C6	0.1567	0.0377	0.3733	0.1155
C7	0.1567	0.0377	0.3146	0.0728
C8*	0.1868	0.0317	NA	NA

Note:
* The single estimate obtained from C8 is a **CMF Estimator derived from frequency-based analysis** (CMF_FBA) from an independent synthetic data set of the same sample size obtained prospectively from the same population of sites.
CMF_PBA = CMF Estimator derived from probability based-analysis

Before analyzing the data, researchers prepared an exploratory analysis consisting mostly of a set of plots for the relationships and trends emerging between the variables of interest. The next section describes the exploratory analysis.

Bias of the OR by True CMF Estimand Value

As discussed in the previous chapter, bias in the OR estimator is expected that depends on the true CMF estimand and the base odds in the population under study. Researchers prepared some graphic assessment of this bias and the performance of the CMF_PBA estimator in correcting for this bias. As a comparison, these plots include the CMF_FBA estimator that should not be biased due to these issues.

The next figures show the performance of the pair of retrospective estimators (OR and CMF_PBA) produced from some of the cases as defined in the prior chapter. The bias for the CMF_FBA is shown in each case for comparison.

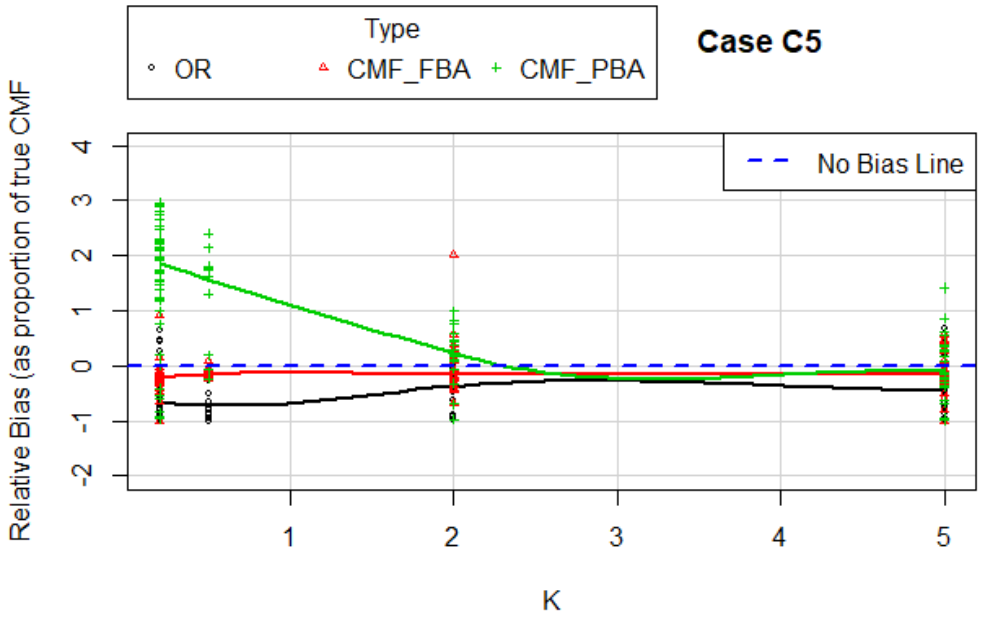


Figure 6. Bias of Estimators from Case 5 by Value of Sampling Rate K, for True CMF=0.25.

Figure 6 shows the bias of the three estimators for Case 5 for samples with at least 200 sites, for the true CMF estimand of 0.25. First, the bias for the CMF_FBA is very small and slightly negative. Second, the bias of the OR remains negative (i.e., underestimation), as expected. Finally, the sampling parameter has a significant impact on the bias of the CMF_PBA. This was expected for all the cases that did not have an adjustment toward the population (Cases 1, 2, and 5). Researchers confirmed that the plots for those cases look almost identical to Figure 6. The bias is not an issue for larger K values, but it exacerbates for smaller k values (i.e., for samples comprising mostly of cases and only few controls in comparison). As said, this was expected because the sample base odds tend away from the population base odds and toward infinity as the k rate decreases.

Researchers confirmed that the biases tend to zero for both retrospective estimators as the true CMF increases, as shown in Figure 7 and Figure 8.

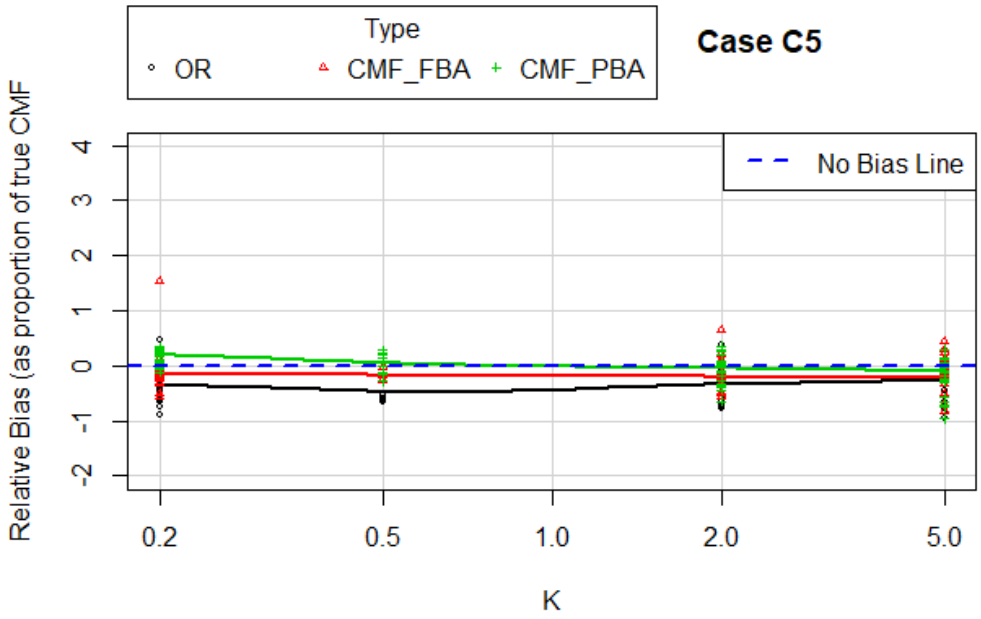


Figure 7. Bias of Estimators from Case 5 by Value of Sampling Rate K, for True CMF=0.75.

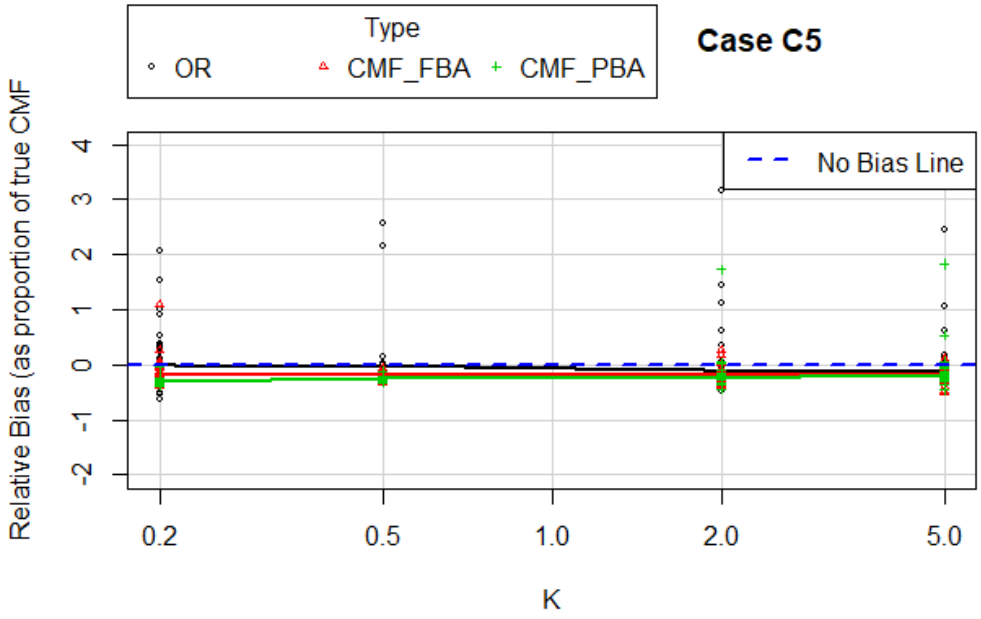


Figure 8. Bias of Estimators from Case 5 by Value of Sampling Rate K, for True CMF=1.5.

Expectedly, the performance of the CMF_PBA is significantly better for the cases where an adjustment of the base odds was done toward the population (Cases 2, 4, 6, and 7), since this correction should have curbed the distortion incorporated by the unbalanced sampling. The

typical performance of the four cases with correction toward the population base odds can be seen in Figure 9. This figure represents the bias for Case 6 estimators when the true CMF is fixed at 0.25. Generally, the improvement in performance was most notable for Cases 6 and 7.

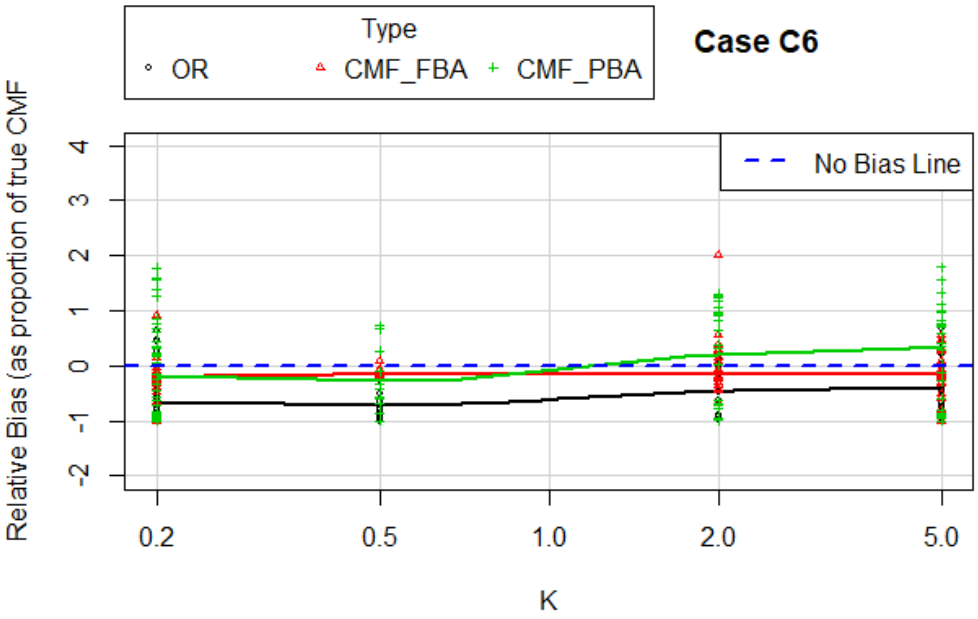


Figure 9. Bias of Estimators from Case 6 by Value of Sampling Rate K, for True CMF=0.25.

The OR estimate is consistently negatively biased, and the CMF_FBA estimator tends to be slightly negatively biased. Interestingly, the bias of the CMF_PBA estimator is on par with the CMF_FBA estimator for the two values of k smaller than 1, and it goes toward a moderately positive bias for k values larger than one. Also interesting is the transition from slight negative bias to slight positive happens approximately at the k value indicated a balanced prospective sample (i.e., k=1).

Although results show that FBA is to be preferred in general, Figure 10 shows that FBA tend to be biased when the crash expectation is lower. In those cases, both estimates from PBA offer virtually unbiased alternatives when the estimand CMF also has small values.

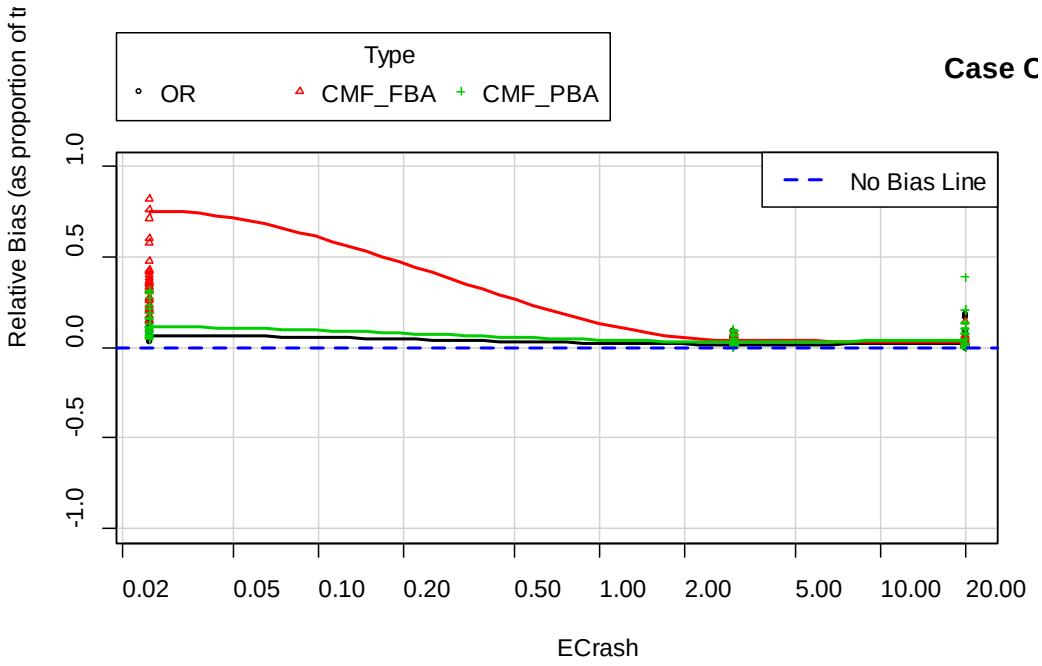


Figure 10. Bias of Estimators from Case 7 by Crash Expectation, for True CMF=0.25.

However, Figure 11 shows that the bias of PBA estimates is expected to go up as the expectation of cashes increases and the CMF under estimation is larger than one. CMF_PBA should then be the preferred method in that case.

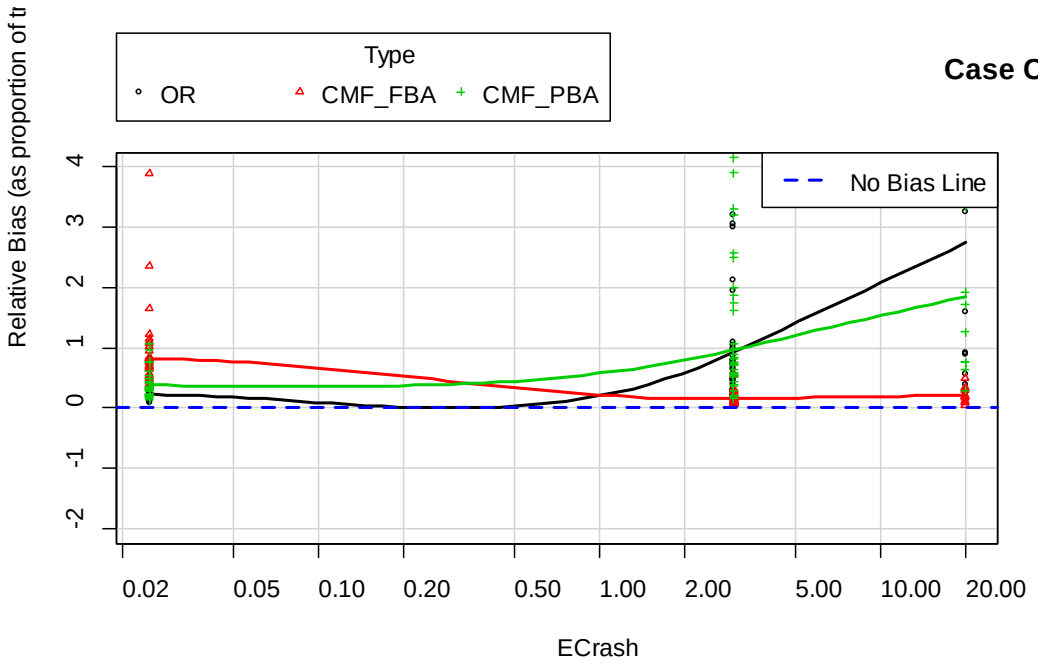


Figure 11. Bias of Estimators from Case 7 by Crash Expectation, for True CMF=1.5.

Proximity Rank and Probability of Capturing True CMF Estimand

Next, researchers looked at marginal comparisons between the three types of estimates obtained from each run. Figure 12 shows the rank in closeness of the three types of estimates among the results (1 being the closest to the true CMF estimand, 15 being the farthest).

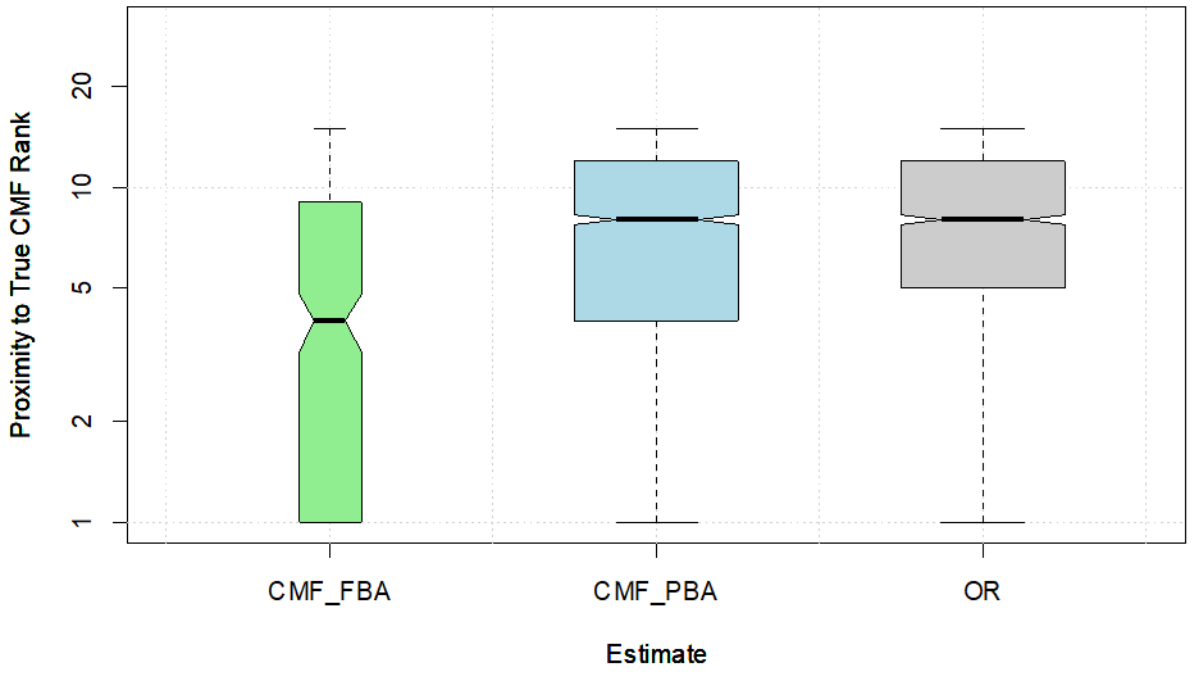


Figure 12. Rank in Proximity to the Parameter for Each Estimate.

This figure shows that the CMF_FBA estimator ranks between first and eighth about 75 percent of the time. In contrast, the CMF_PBA and OR tend to rank very similarly to each other, with perhaps the CMF_ PBA being slightly better, as it tends to rank between fourth and twelfth about 50 percent of the time, while ranking between first and fourth about 25 percent of the time. This is slightly better than the OR that tends to rank between fifth and twelfth about 50 percent of the time, while ranking between first and fifth for 25 percent of the time.

Figure 13 shows that both the OR and CMF_PBA estimators from each case tend to perform similarly in terms of proximity to the true parameter. Judging by the median ranks, the CMF_PBA from cases C2, C4, C6, and C7 —all cases that do some adjustment toward the population from the base odds from the regression model— tend to perform better than their paired OR and both the CMF_PBA and OR from the comparable cases that did not adjust the base odds toward the population (i.e., Cases 1, 3, and 5).

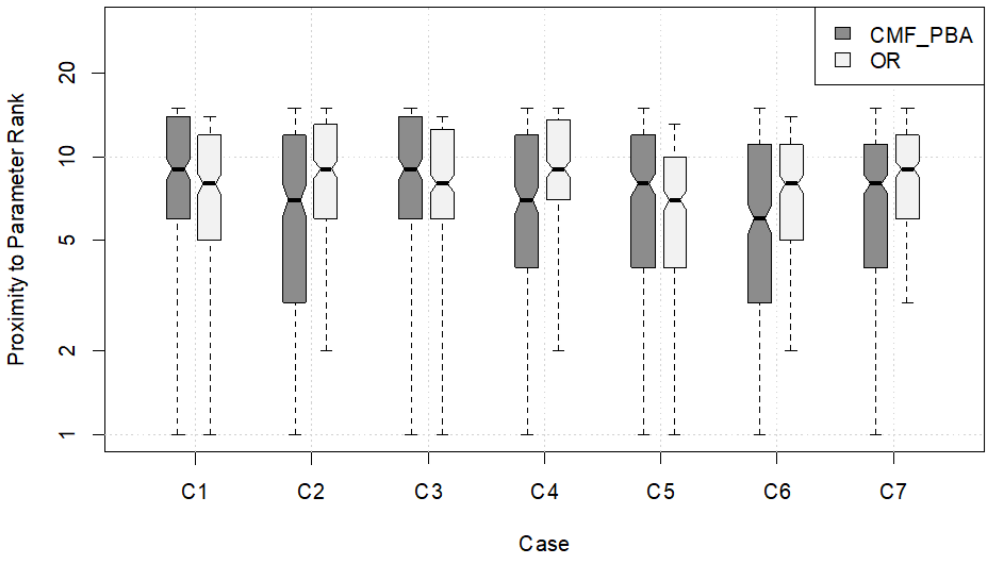


Figure 13. Rank in Proximity to the Parameter by Case and Estimate Type.

Next, researchers calculated a 95 percent CI around each estimate and verified whether the true CMF estimand was captured within that CI. Researchers defined an indicator variable as one if the true estimand was captured, zero otherwise. The expected value of a variable so defined is the probability of capturing the true estimand by the CI. Figure 14 shows the trend line for the probability of success by the true CMF estimand value.

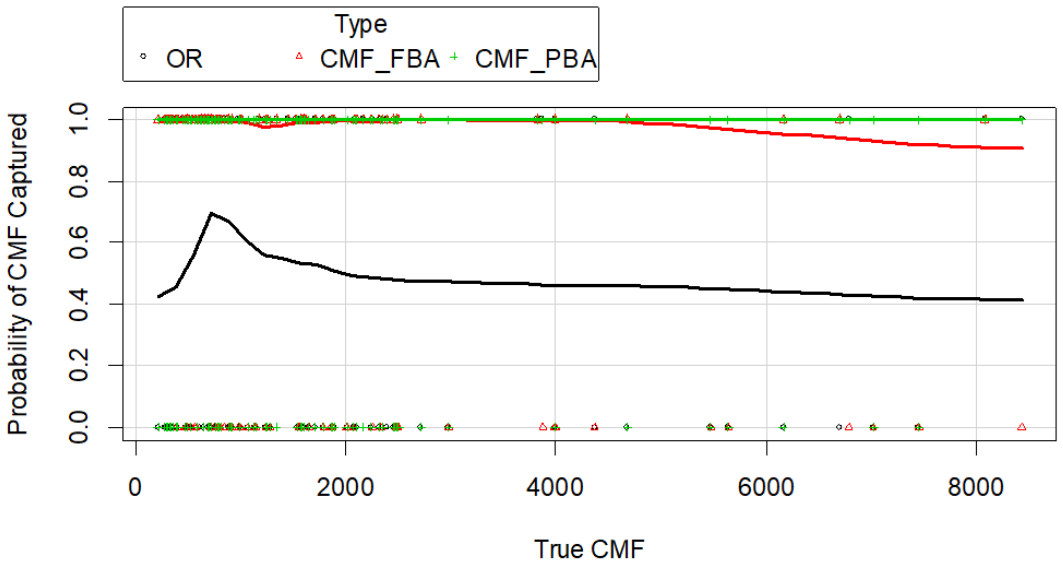


Figure 14. Probability of Capturing True CMF by Estimate Type.

This figure shows the CMF_PBA performs very comparably to the prospective estimator CMF_FBA, and that both outperforming the OR in capturing the true parameter.

Figure 15 shows that the CMF_FBA does not seem to have issues at any particular value of the CMF estimand. In contrast, the two retrospective estimators tend to under-perform at the lower true CMF estimand value (though the CMF_PBA clearly outperforms the OR). This figure also shows that the expected performance of the CMF_PBA improves as the true CMF estimand value increases.

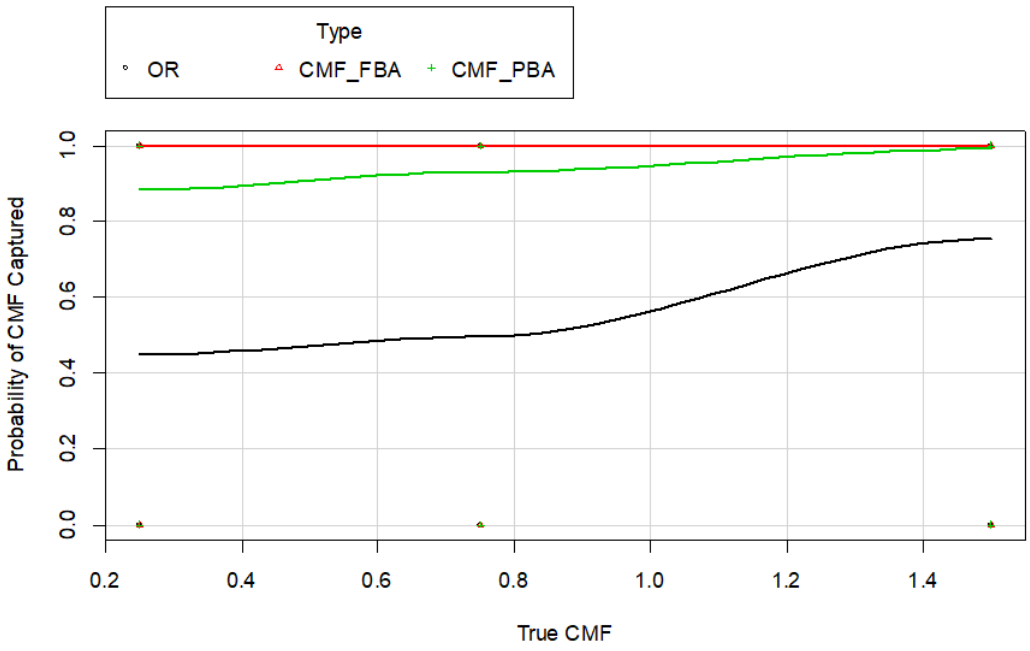


Figure 15. Probability of Capturing True CMF by Estimate Type.

Figure 16 shows that the probability of success of the CMF_PBA estimator is consistently higher than the OR Estimator's across all sample sizes in the experiment.

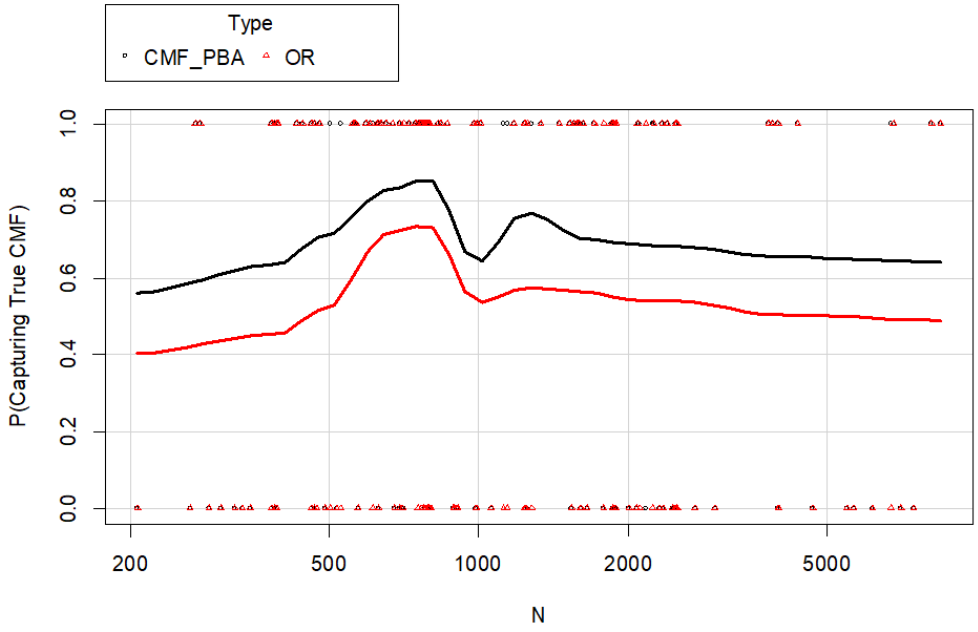


Figure 16. Probability of Capturing True CMF by Sample Size and Estimate Type (Case 7 Only).

In contrast with Figure 16, Figure 17 shows that the proximity of the CMF_PBA estimate to the true parameter is not necessarily higher nor lower than the OR across all sample sizes in the experiment.

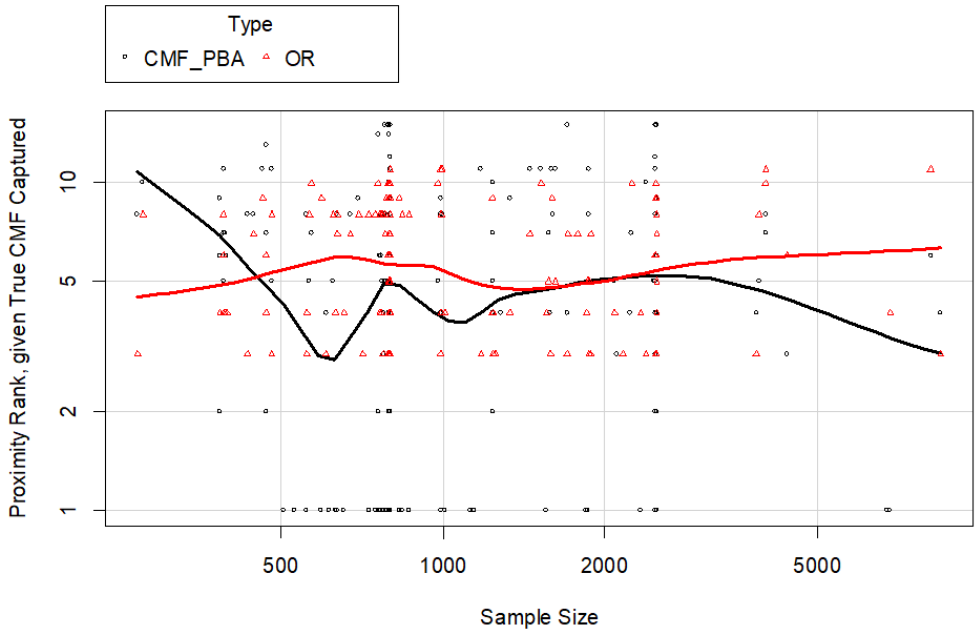


Figure 17. Proximity Ranks by Sample Size, Given True CMF Is Captured (Case 7 Only).

Figure 18 shows how the OR tend to be closer to the true parameter (the true CMF estimand) than the CMF_PBA, given that both estimates have captured the true CMF parameter. This feature suggests a tradeoff of the CMF_PBA estimator—it is more likely to capture the true CMF estimand, but it also tends to be farther away from the actual parameter value.

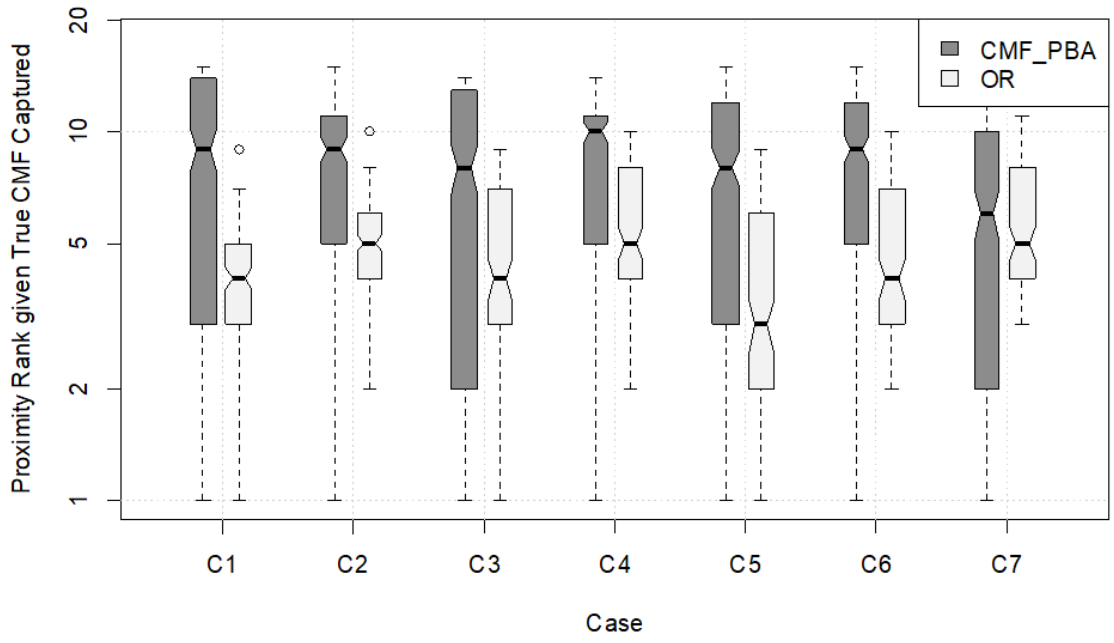


Figure 18. Proximity Rank Given True CMF Was Captured.

Notably, the difference in performance between the OR and CMF_PBA is minimal for the pair of estimates from Case 7 in terms of proximity to the true CMF. Furthermore, Figure 19 shows that the performance of the CMF_PBA is superior for the three levels of the True CMF in the experiment, especially for true CMF = 1.5 (i.e., it had better proximity rankings than the OR on its lower tail).

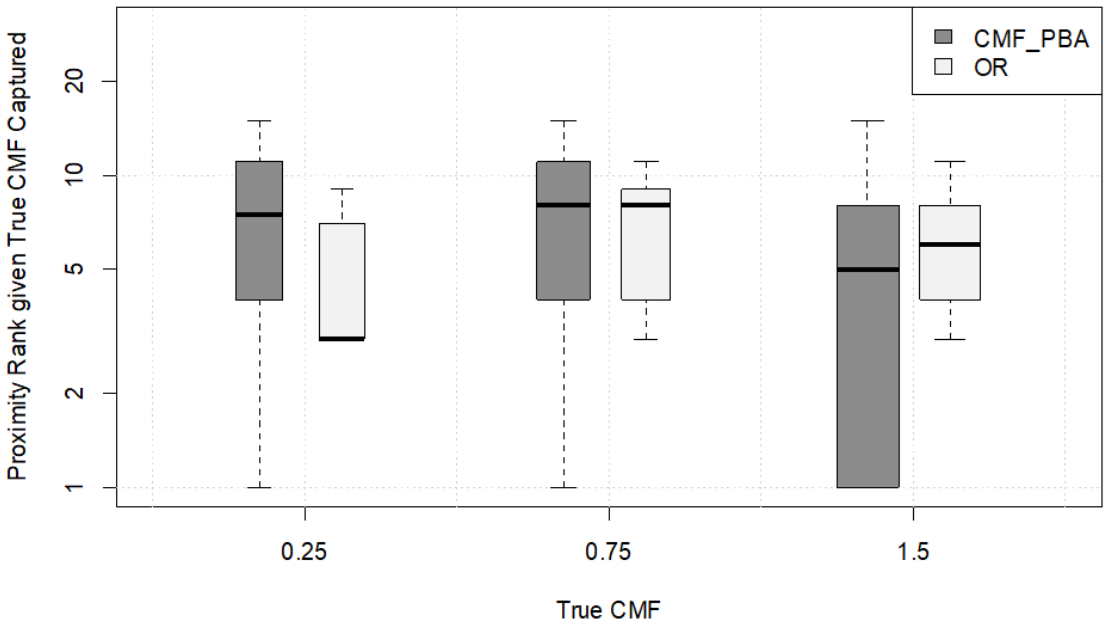


Figure 19. Proximity to True CMF by Estimate Type, Given that CI Contains True CMF (Case 7 Only).

The next section shows the results of a set of formal statistical analyses on the data from the experiment.

Regression Analysis

This section summarizes the results from modeling the three measures of effectiveness as functions of the experiment design parameters. Researchers considered the features observed in the exploratory analyses above to construct appropriate generalized mixed effects models to account for the blocking that results from the fact that each simulation run produced 15 estimates. Researchers performed stepwise model selection based on Akaike information criterion and on Deviance Analysis of Variance in cases of convergence issues.

Probability of Capturing True CMF

The first statistical analysis modeled the probability of capturing the CMF within a 95 percent CI as a function of the experiment parameters. This probability is referred to as the probability of success from this point forward. Table 5 shows the success ORs for the experiment parameters in the most parsimonious model.

Table 5. Analysis Results for the Probability of Capturing the True CMF.

Parameter	OR (95 % CI)		Significance[a]
	Lower Limit	Upper Limit	
True CMF Estimand	1.7515	9.9923	**
Sampling rate (k)	0.4786	0.8118	***
Number of cases	0.9978	1.0000	*
Parameters that Affected Performance of the OR Estimator			
Base Odds for OR	0.1711	0.6180	***
Theta (for OR)	0.9788	0.9994	*
E(Crashes) (for OR)	0.8294	0.9433	***
Population Adjustment (for OR)	0.7587	1.4081	
Comparative Effects on the CMF_PBA Estimator			
Theta (for PBA)	0.9866	1.0076	
E(Crashes) (for PBA)	0.9348	1.0650	
Population Adjustment (for PBA)	0.5350	1.0099	

[a] Significance Levels are as follows:

* = Statistically different from 0.0 at the 5.0% significance level.

** = Statistically different from 0.0 at the 1.0% significance level.

*** = Statistically different from 0.0 at the 0.1% significance level.

The next subsections describe the implications of the results shown in Table 5.

Experiment Parameters that Affected the Performance of both Estimators Equally

Table 5 shows that three design parameters had a uniform effect on both the CMF_PBA and the OR estimators: 1) the value of the true CMF; 2) the sampling rate k; and 3) the number of cases (as opposed to the number of controls).

The odds of success for both estimates (CMF_PBA and OR) increased by a factor of between 1.75 and 9.99 for each additional point in the CMF. Alternatively, results indicate that the odds of successfully capturing the true CMF increased by a factor of between 1.05 and 1.26 for each 0.1 increase in the CMF to be estimated.

The odds of success for both estimates (CMF_PBA and OR) decreased significantly by a factor of between 0.479 and 0.812 for each unit increase in the sampling rate k.

The number of cases (i.e., locations with at least one crash) had a mild effect on the odds of success of both estimators. For each additional case in the sample, the odds of success decreased by a factor of between 0.998 and 1.000. Alternatively, results indicate that the odds of successfully capturing the true CMF decreased by a factor of between 0.799 and 1.000 for each additional 100 cases in the sample.

Table 5 shows two sets of estimates for the three experimental parameters Theta, crash expectation, and the variable indicating an adjustment for the population on the base odds in the CMF_PBA estimator. The first set of estimates reflects the impact of these parameters on the OR estimator and the second set the corresponding estimates for the CMF_PBA estimator. There is no evidence of any effect on the odds of successfully capturing the true CMF (at a 95 percent confidence level).

Experiment Parameters that Affected the Performance of the OR Estimator

The success rate of the OR estimator was found to suffer at three levels in addition to the effects found in common for both the OR and CMF_PBA estimators. The second set of effect estimates in Table 5 shows two factors that affected the performance of the OR estimator only. The estimate for the effect of the population adjustment is also shown, but it is not statistically significantly different than 1.0.

After accounting for all other influential factors, the base odds of success for the OR estimator were smaller by a factor of between 0.171 and 0.618 compared to the base odds of the CMF_PBA estimator. In other words, the CMF_PBA was found to be between 1.6 to 5.8 times more likely to succeed in capturing the true CMF, after discounting the effect of design variables.

The inverse dispersion parameter Theta had a significant effect on the odds of success of the OR estimator, in contrast to no discernible effect on the odds of success of the CMF_PBA estimator, as described in the next section. For each unit increase in Theta, the odds of success for the OR estimator decreased by a factor of between 0.9788 and 0.994. Alternatively, results indicate that the odds of successful capture of the true CMF for the OR estimator decreased by a factor of between 0.807 and 1.000 for each 10-units increase in the inverse dispersion parameter Theta.

The expectation of crashes also had a significant effect on the odds of success for the OR estimator, in contrast to no discernible effect on the odds of success for the CMF_PBA estimator (as described in the next section). For each unit increase in crash expectation, the odds of success decreased by a factor of between 0.0829 and 0.943 for the OR estimator.

Experiment Parameters that Affected the Performance of the CMF_PBA Estimator

Table 5 shows the estimated effects on the CMF_PBA estimator of the two experiment parameters found to affect OR performance (Theta and crash expectation) and an estimate for the variable indicating if the estimator includes an adjustment toward the population base odds. The analysis found no evidence supporting the hypothesis that any of these two estimates affected the chances of the CMF_PBA estimator successfully to capture the true CMF (at a 95 percent confidence level).

Comparative Analysis of Standard Errors of the Estimators

In this analysis, researchers calculated the ratio between the SEs of all retrospective estimators to the SE of the prospective estimator (i.e., Case 8). This ratio is called the SE inflation factor (SEIF) from this point forward. The purpose of the analysis is to assess the magnitude of the tradeoff of using the CMF_PBA estimator. On the one hand, the prior section demonstrated that the CMF_PBA is in general more likely to succeed in capturing the true CMF, as well as more robust against the detrimental effects of increasing crash mean and the dispersion parameter.

Researchers fitted a linear mixed-effects multiplicative model to characterize the median SEIF as function of the experiment parameters. Table 6 shows the results after a stepwise model selection process. This table is divided into two sets of parameters.

Table 6. Analysis Results for the SEIF on the Retrospective Estimators.

Parameter	Estimate	Std. Error	Degr. of freedom	t value	Pr(>|t|)	Significance[a]
True CMF Estimand	1.13E+00	3.32E-01	4.09E+02	3.38E+00	7.86E-04	***
Crash Expectation	4.82E-02	2.17E-02	4.09E+02	2.23E+00	2.64E-02	*
Sampling rate (k)	-1.55E-01	8.93E-02	4.09E+02	-1.73E+00	8.36E-02	#
Number of Cases in Sample	3.10E-04	1.85E-04	4.09E+02	1.67E+00	9.54E-02	#
Estimator Magnitude	6.07E-08	1.63E-08	5.48E+03	3.73E+00	1.96E-04	***
Estimator SE	1.87E-13	1.21E-14	5.40E+03	1.54E+01	< 2e-16	***
Basel Inflation OR SE	-4.49E-01	4.28E-01	4.11E+02	-1.05E+00	2.94E-01	
Base Inflation CMF_PBA SE	2.49E+00	6.77E-02	5.39E+03	3.68E+01	< 2e-16	***
CMF_PBA from Model 5,6 and 7	-8.14E-01	1.04E-01	5.39E+03	-7.87E+00	4.38E-15	***
CMF_PBA from Model 1_2 and Population Adjustment	-1.61E+00	1.03E-01	5.39E+03	-1.55E+01	< 2e-16	***
CMF_PBA from Model 3_4 and Population Adjustment	-1.61E+00	1.03E-01	5.39E+03	-1.55E+01	< 2e-16	***

[a] Significance Levels as follows:
= Statistically different from 0.0 at the 10.0% significance level.
* = Statistically different from 0.0 at the 5.0% significance level.
** = Statistically different from 0.0 at the 1.0% significance level.
*** = Statistically different from 0.0 at the 0.1% significance level.

The upper half corresponds to the effects of experiment parameters that affected the performance of both estimators (OR and CMF_PBA). The lower half corresponds to the differentiated effects of some parameters over the OR and CMF_PBA separately.

Parameters that Affect Both Retrospective Estimators (OR and CMF_PBA)

Other things equal, Table 6 indicates that higher values of the true CMF under investigation resulted in an inflation of 3.08 (3.08=exp(1.13)) on the SEs of both retrospective estimators for the prospective estimator from a similar sample size, for each increase of 1 in the estimand CMF. Alternatively, it is expected that the SE of a retrospective estimate will increase by a factor of 1.76 for each additional 0.5 increase in the magnitude of the true CMF under estimation (1.76=exp (1.12*0.5)).

Increases in the crash expectation were also found to result in inflation of the SEs for both retrospective estimators. After controlling for other factors, retrospective SEs were found to increase by a factor of 1.05 (1.05=exp (4.82×10^{-2})) for each one crash increase in the grand average of the crash expectation in the sample under study. Results also imply that the median SEIF increases by a factor of 1.62 (1.62=exp (10x 4.82×10^{-2})) for each 10 crashes increase in the grand average of crash expectation at the sample of sites under study.

This analysis did not find significant evidence that the sampling rate and the number of cases in the sample affect the inflation of SEs of retrospective estimators (at a 95 percent confidence level). However, at a 90 percent confidence level, the evidence is suggestive that decreasing sampling rate value and increasing number of cases in the sample could each result in median SEIF inflation.

This analysis found significant evidence that the amount of median SEIF is directly proportional to the magnitude of the retrospective estimator and its SE. The rate of increase of these effects is very small and of no practical significance to this research.

Specific Effects on the CMF_PBA Estimator and Relative SEIF between OR and CMF_PBA

As mentioned earlier, the lower half of Table 6 has estimates that represent variability in SEIF of specific retrospective estimators. For easier interpretation, researchers used those parameters to construct the ratios and their CIs shown in Table 7.

Table 7. Analysis Results for the SEIF on the Retrospective Estimators.

Median SEIF Ratios	Estimate	Std. Err.	95% CI	
			Lower	Upper
OR / CMF_FBA	0.638	0.273	0.276	1.476
CMF_PBA / OR (CMF_PBA from Cases 1 or 2)	2.430	0.219	2.036	2.901
CMF_PBA / OR (CMF_PBA from Cases 3 or 4)	2.430	0.219	2.036	2.901
CMF_PBA / OR (CMF_PBA from Cases 5, or 6)	5.359	0.484	4.489	6.397
CMF_PBA / OR (CMF_PBA from Case 7)	1.763	0.119	1.544	2.014

The first ratio in Table 7 indicates that the median SEIF for the OR estimator is between 0.276 and 1.476, after accounting for other factors. In other words, there is no evidence that the SE of the OR estimator and the prospective estimator (i.e., the CMF FBA) are statistically different, after accounting for other factors influential to the SEIF of the retrospective estimators.

In contrast with the previous finding, the last three ratios in Table 7 show that the SE of the CMF_PBA estimator is, in general, wider than the SE of the OR. This finding explains, to a large extent, why the CMF_PBA performed significantly better than the OR in capturing the true CMF in a 95 percent CI. Larger SEs result in wider CIs and increased robustness against influential factors.

Table 7 indicates that, other things equal, the median SE for the CMF_PBA is between 2.036 and 2.91 times as wide as the SE for the OR when the CMF_PBA was estimated from cases 1 through 4, with little to no difference between these cases.

Finally, Table 7 indicates that the median SE for the CMF_PBA obtained from Case 7 is between 1.544 and 2.014 times as wide as the SE for the OR. Researchers consider that such range for SEIF is an acceptable price to pay for having the odds of capturing the true CMF in a CI increased significantly (odds increase by a factor of between 1.5 and 5.8, per Table 5).

Summary

This chapter has summarized the analyses performed between the two retrospective estimators for a CMF that concern this report. Some important points are summarized next:

- Researchers coded the crash model for two-way, two-lane model in the HSM to generate synthetic samples for analysis.
- Researchers adopted a matching sampling scheme because of two reasons: 1) the inclusion of AADT in the crash simulation process, a variable that significantly affects the

number of crashes generated; and 2) AADT is expected to affect the base odds in the model significantly, now conditional to AADT and to the OR. A matched sample scheme allowed researchers to explore various ways account for the AADT in the estimation of the CMF_PBA estimators.

- Eight cases were proposed to construct the CMF_PBA estimator based on three specifications of the logistic model to be fitted on the retrospective samples analyzed.
- Researchers proposed an experiment design including six multilevel factors with a total of 486 potential combinations. Because researchers planned to develop 15 estimates for each of those combinations of factors, researchers set an initial target of 7,290 estimates.
- After running the simulations, it was not possible to generate synthetic data sets for all 486 possible combinations of factors. As 71 combinations did not yield synthetic data, researchers generated a total of 6,225 estimators for the analysis.
- Researchers established two criteria for evaluation of the retrospective estimators: 1) the probability to capture the true CMF estimand; and 2) SEIF with respect to a comparable prospective estimator (i.e., CMF_FBA).
- An exploratory analysis of the retrospective estimates showed that:
    - The prospective estimator CMF_FBA tends to fall closer to the true CMF estimand.
    - The CMF_FBA estimator tends to outperform the OR in capturing the true CMF estimand, but the CMF_PBA tends to outperform both the CMF_FBA and the OR.
    - The value from the CMF_PBA estimator tends to be closer than the OR estimator to the true CMF estimand for all cases that did an adjustment of the regression base odds toward the population base odds when constructing the CMF_PBA.
    - Among all estimates that captured the true CMF estimand in a 95 percent CI, the OR tended to be closer to the estimand, except for Case 7, where the CMF_PBA outperformed the OR.
- A formal statistical analysis on the probability of success in capturing the estimand CMF yielded the following results:
    - Increasing the estimand CMF, sampling rate, and number of sites-with crashes in the sample will result in increased probability of capturing the true CMF estimand for both retrospective estimators, ceteris paribus.
    - In general, the probability of success is smaller for the OR compared to either the CMF_PBA or the CMF_FBA estimators. The probability of success for the OR decreases with increasing dispersion and increasing number of expected crashes. The CMF_PBA was robust against these adverse effects of increased dispersion and crash expectation.

- Researchers performed a formal statistical analysis on the inflation of the SE for the retrospective estimators with respect to the SE for the comparable prospective estimator. The results indicate that:
    - The SE width increases for both the CMF_PBA and OR when the CMF estimand, crash expectation, and the number of sites with crashes in the sample increase. Conversely, the SEIF increases and when the sampling rate decreases.
    - The median SE width of the OR is comparable to the CMF_FBA SE. In general, the median SE for all CMF_PBA estimators is larger than the median SE for the OR. The best performing CMF_PBA was the one obtained from Case 7. All CMF_PBA estimates from Cases 1 through 5 performed comparably well but not as well as the estimate from Case 7. The worst performing CMF_PBA estimators were those from Cases 5 and 6.

CHAPTER 5. CONCLUSIONS AND FUTURE DIRECTIONS

The motivation for this work was to provide guidance to future researchers when estimates of safety effects are needed for crash types that are especially scarce. In these cases, traditional analysis methods are impractical to implement.

This project identified and quantified relationships between the OR from PBA and the CMF from crash FBA. Researchers developed an analytical framework in which the crash generating process gives emergence to the link between the OR and the CMF in that context.

Researchers proposed a set of adjustments to curb the known bias of OR in estimating the CMF from the analytical framework. The set of adjustments were evaluated using synthetic data sets that researchers generated for that purpose.

The following sections summarize the findings, conclusions, and recommendations from this work

On the Bias of the OR and the Correction Offered by the CMF_PBA

As it was uncovered in the exploratory analysis of the results, the behavior of the OR estimator confirmed a systematic bias in estimating the true CMF as researchers exposed in Chapter 3 by analytical means (i.e., theorem 3.1 expressed in Equation 13 or Equation 14). Furthermore, the analysis of simulated data found severe biases in the CMF_PBA estimator from the cases that assumed that the base odds from the sample could be taken to represent the base odds from the population (i.e., applying Equation 13 without adjusting the base odds toward the population, as required by Equation 14). However, the CMF_PBA was essentially unbiased (or exhibited bias comparable to the bench mark CMF_FBA estimator) when the appropriate adjustment is performed on the base odds prior to calculations. In Cases 2, and 4, this adjustment consisted of multiplying the base odds by the sampling rate, as shown in Equation 13. Additionally, Equation 14, but researchers considered other ways to perform an equivalent adjustment with similar or perhaps slightly better results. In Case 6, the adjustment is a weighted linear combination of coefficient estimates weighted toward mean values of the exposure variable (AADT) to produce an estimate of the population base odds equivalent to an estimate of the population base odds from the marginal distribution of crashes in the sample. With similar results, the adjustment in Case 7 was simply the combination of the OR from the regression with a population base odds estimate from marginal base odds of the sample and the sampling rate, per Equation 14.

It is recommended that researchers interested in applying PBA consider either of the two formulations of theorem 3.1 (Equation 13 or Equation 14) to perform the adjustment to the CMF_PBA estimator to correct the known bias of the OR estimator with respect to the true

CMF. This adjustment is especially needed when the estimand CMF tend to have smaller values, as the OR tends be consistently below the true CMF, which implies that it would offer an over-optimistic estimate of effectiveness for CMFs smaller than one.

On the Viability of Estimating CMFs from Retrospective Analyses

The analyses found that it is generally preferable to perform a prospective analysis to estimate the CMF, as this estimate is least biased, it offers a high probability of capturing the true CMF estimand and offers relatively narrow CIs. The first two of these advantages are shared with the CMF_PBA with population adjustment in most cases, but at the expense of a wider SE (between 1.54 and 2.01 times as large, per the SEIF analysis).

Although results indicated that FBA should be preferred in general, this research showed that CMF_FBA has a significant bias in cases where the crash expectation is small. PBA estimates offer an unbiased alternative to the CMF_FBA across a wide range of crash expectations when estimating small CMFs (Figure 10). Researchers recommend the CMF_PBA in those cases. However, the analyses found the CMF_FBA would have positive bias when estimating larger CMFs when the crash expectations are higher (Figure 11). Fortunately, the bias from the CMF_FBA obtained from traditional methods is minimal in those cases, so it should be preferred. Regardless, Figure 11 shows that the CMF_PBA tend to correct for the OR bias at high crash expectations (yet not completely eliminating the bias).

On the Performance of CMF_PBA Estimator with Respect to the OR

In general, the CMF_PBA estimator was found more robust and less biased than the OR in terms of higher probability of success in capturing the true CMF parameter in general (because of the significantly lower base odds of success for the OR, per Table 5). Additionally, the CMF_PBA was not found sensitive to the dispersion and expected value of crashes, both factors found to significantly affect the OR probability of success in capturing the true CMF parameter (per Table 5). However, researchers found the CMF_PBA to be less biased and robust against increasing dispersion and crash expectation, but at the price of a larger SE than the SEs for the OR estimators, which implies that it would require larger samples in general (per Table 7).

Recommendations for Future Researchers

Among the different CMF_PBA with adjustment toward the population, the CMF_PBA estimator in Case 7 was found especially robust and is recommended as the best alternative to the CMF_FBA estimator when a prospective sample is not feasible to obtain. The CMF_PBA from Case 7 is obtained from a retrospective model (i.e., logistic regression) that includes a term for AADT. The PBA estimator is constructed with the OR from the model results and an estimate of the population base odds that combines the sampling rate and a marginal estimate

of the base odds in the sample. When constructing the SE for this estimator, a critical assumption is that the OR and the adjusted base odds are independent, though these two quantities are probably correlated (most likely negatively). The reasoning behind the assumption is that the estimated SE is wider than the true SE if the correlation is indeed negative.

Future Work

Researchers recommend future work on the following points:

- Further refinements and theoretical characterizations for the CMF_PBA estimator and simple adjustments to the base odds, similar to the adjustment proposed in Case 7.
- Develop a software implementation to estimate robust CMF_PBA estimators.
- Determine the set of conditions (e.g., required sample size, crash expectation, dispersion, and true CMF estimand) under which the CMF_PBA should be used over the traditional estimation methods.

REFERENCES

[1] F. L. Ramsey and D. W. Schafer, The Statistical Sleuth. A Course in Methods of Data Analysis. Second Edition, Pacific Grove, CA: Duxbury, 2002.

[2] AASHTO, Highway Safety Manual, Washington, D.C.: Transportation Research Board of National Academies, 2010.

[3] F. Gross and P. P. Jovanis, "Estimation of Safety Effectiveness of Lane and Shoulder Width: Case-Control Approach," *Journal of Transportation Engineering Vol. 133,* pp. pp. 362-369, 2007.

[4] C. Buth, W. Williams, M. Brackin, D. Lord, D. Geedipally and A. Abu-Odeh, "Analysis of Large Truck Collisions with Bridge Piers," Texas Department of Transportation, Austin, TX, 2010.

[5] R. Avelar, M. Pratt, J. Miles, N. Trout and J. Crawford, "Develop Metrics of Tire Debris on Texas Highways," Texas Department of Transportation, Austin, TX, 2016.

[6] F. Gross, B. Persaud and C. Lyon, "A Guide to Developing Quality Crash Modification Factors," FHWA, Washington, DC, 2010.

[7] D. Carter, R. Srinivasan, F. Gross and F. Council, "Recommended Protocols for Developing Crash Modification Factors," American Association of State Highway and Transportation Officials (AASHTO), 2012.

[8] AASHTO, Highway Safety Manual, Washington, DC: American Association of State Highway and Transportation Officials, 2010.

[9] F. Gross, B. Persaud and C. Lyon, "A Guide to Developing Quality Crash Modification Factors," Federal Highway Administration (FHWA), Washington, DC, 2010.

[10] F. Gross and E. T. Donnell, "Case–control and cross-sectional methods for estimating crash modification factors: Comparisons from roadway lighting and lane and shoulder width safety effect studies," *Journal of Safety Research,* vol. 42, pp. 117-129, 2011.

[11] E. J. A. B. G. M. Hauer, "Screening the Road Network for Sites With Promise," *Transportation Research Record: Journal of Transportation Research Board,* vol. 1784, pp. 27-32, 2002.

[12] J. Park and M. Abdel-Aty, "An Alternative Approach for Combining Multiple Crash Modification Factors Using Adjustment Function and Analytic Hierarchy Process," in *TRB 96th Annual Meeting Compendium of Papers*, Washington D.C., 2017.

[13] E. Hauer, Observational Before-After Studies in Road Safety, Oxford: Elsevier Science Ltd, 1997.

[14] G. Bahar, "Methodology for the Development and Inclusion of Crash Modification Factors in the First Edition of the Highway Safety Manual," *Transportation Research Circular,* 2010.

[15] T. Saleem and A. Lorion, "Estimating Crash Modification Factors Using Surrogate Measures of Safety," in *25th CARSP Conference*, Ottawa, Ontario, 2015.

[16] J. M. Violanti and J. R. Marshall, "Cellular Phones and Traffic Accidents: An Epidemiological Approach," *Accident Analysis and Prevention,* vol. 28, no. 2, pp. 265-270, 1996.

[17] O. Drummer, J. Gerostamoulos, H. Batziris, M. Chu, J. Caplehorn, M. Robertson and P. Swann, "The involvement of drugs in drivers of motor vehicles," *Accident Analysis and Prevention,* vol. 36, pp. 239-248, 2004.

[18] A. Filtness, K. Armstrong, A. Warson and S. Smith, "Sleep-Related Vehicle Crashes on Low Speed Roads," *Accident Analysis and Prevention,* vol. 99, pp. 279-286, 2017.

[19] H. Leitgöb, "The Problem of Modeling Rare Events in ML-based Logistic Regression," in *The fifth Conference of the European Survey Research Association (ESRA)*, Ljubljana, Slovenia, 2013.

[20] S. S. J. Gao, "Asymptotic Properties of a Double Penalized Maximum Likelihood Estimator in Logistic Regression," *Statistics and Probability Letters,* pp. 925-930, 2007.

[21] T. Ghosh, "Logistic Regression with Low Event Rate," 23 June 2013. [Online]. Available: https://www.slideshare.net/tejamoy/logistic-regression-with-low-event-rate-rare-eve.

[22] D. Firth, "Bias Reduction of Maximum Likelihood Estimates," *Biometrika,* vol. 80, pp. 27-38, 1993.

[23] G. Heinze and M. Schemper, "A Solution to the Problem of Separation in Logistic Regression," *Statistics in Medicine,* vol. 21, pp. 2409-2419, 2002.

[24] P. Peduzzi, J. Concato, E. Kemper, H. T. R and A. R. Feinstein, "A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis," *Journal of Clinical Epidemiology,* vol. 49, pp. 1373-1379, 1996.

[25] S. Tanish, "Framework to Build Logistic Regression Model in a Rare Event Population," 18 January 2014. [Online]. Available: https://www.analyticsvidhya.com/blog/2014/01/logistic-regression-r.

[26] S. Das, X. Sun, F. Wang and C. Leboeuf, "Estimating Likelihood of Future Crashes for Crash-Prone Drivers," *Journal of Traffic and Transportation Engineering,* vol. 2, no. 3, pp. 145-157, 2015.

[27] A. Theofilatos, G. Yannis, P. Kopelias and F. Papadimitriou, "Predicting road accidents: a rare-events modeling approach," *Transportation Research Procedia,* vol. 14, p. 3399 – 3405, 2016.

[28] G. King and L. Zeng, "Logistic Regression in Rare Events Data," 6 2 2001. [Online]. Available: https://gking.harvard.edu/files/0s.pdf.

[29] G. King and L. Zeng, "Logistic regression in rare events data," 6 2 2001. [Online]. Available: https://gking.harvard.edu/files/0s.pdf.

[30] M. L. Veazey, E. .. C. Franklin, C. Kelley, J. Rooney, L. Frazer and R. J. Toonen, "The implementation of rare events logistic regression to predict the distribution of mesophotic hard corals across the main Hawaiian Islands," *Peer J,* 2016.

[31] C. O. Schmidt and T. Kohlmann, "When to use the odds ratio or the relative risk?," *International Journal of Public Health,* vol. 53, pp. 165-167, 2008.

[32] D. Banks, B. Persaud, C. Lyon, K. Eccles and S. Himes, "Enhancing Statistical Methodologies for Highway Safety Research," FHWA, 2014.

[33] P. P. Shrestha and K. J. Shrestha, "Factors associated with crash severities in built-up areas along rural highways of Nevada: A case study of 11 towns," *journal of traffic and transportation engineering,* vol. 4, no. 1, pp. 96-102, 2017.

[34] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks and D. J. Ramsey, "The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-car Naturalistic Driving Study Data," 2006.

[35] R. Young, "Naturalistic Studies of Driver Distraction: Effects on Analysis Methods on Odds Ratios and Population Attributable Risk," in *Seventh International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, 2013.

[36] M. Abdel-Aty, N. Uddin, A. Pande, M. Abdalla and L. Hsia, "Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression," *Transportation Research Record,* vol. 1897, pp. 88-95, 2004.

[37] R. Yu, X. Wang and M. Abdel-Aty, "A Hybrid Latent Class Analysis Modeling Approach to Analyze Urban Expressway Crash Risk," *Accident Analysis and Prevention,* vol. 101, pp. 37-43, 2017.

[38] L. C. D. L. J. S. S. Lanza., "PROC LCA: A SAS Procedure for Latent Class Analysis," *Structural Equation Modeling: A Multidisciplinary Journal,* vol. 14, no. 4, pp. 671-694, 2007.

[39] J. P. Gooch, V. V. Gayah and E. T. Donnell., "Quantifying the Safety Effects of Horizontal Curves on Two-Way, Two-Lane Rural Roads," *Accident Analysis & Prevention,* vol. 92, pp. 71-81, 2016.

[40] L. Wu, D. Lord and S. Geedipally, "Developing Crash Modification Factors for Horizontal Curves on Rural Two-Lane Undivided Highways using a Cross-Sectional Study," pp. Paper No.: 17-02498, November 2016.

[41] W. J. H. O. Bartlett and R. J. Carpenter, "Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression," *American Journal of Epidemiology,* vol. 182, no. 8, pp. 730-736, 2015.

[42] L. R. H. Lin. J ., "Accounting for Informatively Missing Data in Logistic Regression by Means of Reassessment Sampling," *Statistics in Medicine,* vol. 34, no. 11, pp. 1925-1939, 2015.

[43] D. Lord, S. Washington and J. Ivan, "Poisson, Poisson-gamma, and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory," *Accident Analysis and Prevention,* no. 37, pp. 35-46, 2005.

[44] G. H. E. Bahar, "User's Guide to Develop Highway Safety Manual Safety Performance Function Calibration Factors," National Cooperative Highway Research Program, 2014.

[45] K. Dixon, R. Avelar and C. Monsere, "Oregon Signalized Intersection Safety Performance Functions and the Effect of Speed," in *TRB 95th Annual Meeting Compendium of Papers*, Washington, D.C, 2016.

[46] G. Nieto, *Development of Safety Performance Functions (SPF) and Crash Modification Factors (CMF) for Rural Local Roads in Alabama,* University of South Alabama., 2017.

[47] P. Savolainen, T. Gates, D. Lord, S. Geedipally, E. Rista, T. Barrette, R. B. and R. Hamzeie, "Michigan Urban Trunk Line Intersections Safety Performance Functions Development and Support," Michigan Department of Transportation, Lansing, MI, 2015.

[48] R. Srinivasan, D. Carter and K. Bauer, "How to Choose between Calibration SPFs from the HSM and Developing Jurisdiction-Specific SPFs," Federal Highway Administration, 2013.

[49] R. Srinivasan, D. Carter and K. Bauer, "Safety Performance Function Decision Guide: SPF Calibration vs SPF Development," Federal Highway Administration, Washington, USA,, 2013.