

Applying Empirical Bayes Methods and Machine Learning to Network-Level Road Safety Screening

Abdulkareem Olakunle Afolabi [1*], Mercy Osato Omoifo-Irefo [2]
[1] Department of Transportation and Highway Engineering, South Carolina State University, South Carolina, USA
[2] Department of Environmental and Public Health Sciences, University of Cincinnati. Ohio, USA

* Corresponding Author: **Abdulkareem Olakunle Afolabi**

Article Info

E-ISSN: 3107-7099
Volume: 02
Issue: 06
Received: 06-09-2026
Accepted: 04-10-2026
Published: 02-11-2026
Page No: 01-12

Abstract
Reliable network screening requires estimates that distinguish persistent crash burden from random variation and support decisions under limited investigation capacity. This study evaluates whether flexible machine-learning prior means improve Empirical Bayes screening beyond a conventional negative-binomial benchmark. Publicly accessible Montana crash records for 2019–2023 were linked to a fixed 2019 traffic-segment inventory. The resulting network contained 6,169 segments and 78,560 matched crash records. Model development used 2019–2020 counts, while 2021 counts supplied an independent site-history update. Entire corridors were separated into development, calibration and test groups. Negative-binomial regression, Poisson-criterion random forests and Poisson-loss gradient boosting supplied prior means. A Poisson–gamma model, with dispersion estimated on separate calibration corridors, defined the EB update. The untouched evaluation comprised 1138 test segments and 3,888 crashes during 2022–2023. Boosting reduced standalone root mean squared error from 4.521 to 4.419 crashes per segment over two years. Conventional EB nevertheless achieved lower mean absolute error than boosting-assisted EB, 1.875 versus 1.887. Selecting the highest-ranked 10% of test sites captured 50.95% and 50.75% of future crashes, respectively. The paired corridor-bootstrap difference for boosting-assisted EB was −0.21 percentage points, with a 95% interval from −0.68 to 0.85. These findings support EB updating but provide no evidence that the tested machine-learning integration improves conventional EB screening on this network. Interpretation is restricted to reported crash frequency under fixed baseline exposure. Severity information, contemporary exposure and external-network validation remain necessary for broader deployment.

DOI: https://doi.org/10.54660/IJRCE.2026.2.6.01-12

Keywords: Empirical Bayes, machine learning, network screening, road safety, safety performance functions, crash prediction, hotspot identification, potential for safety improvement

1. Introduction

1.1. Background and Practical Motivation

Road agencies must decide which locations warrant detailed safety investigation before they can design and fund appropriate countermeasures. Screening therefore has an operational purpose: to allocate investigative attention across a network using a clearly defined measure of safety performance. A useful procedure should identify locations with substantial expected crash burden while making the consequences of limited data and limited resources visible to decision makers [4].

Observed crash totals are useful records of past events but imperfect estimates of underlying safety. A location with several crashes in one year may experience fewer in the next without any intervention. Conversely, a low observed count does not

establish a low long-run risk. EB estimation addresses this instability by combining information from comparable sites with a location's own crash history [2].

Exposure also matters. Two segments with the same crash count may differ greatly in traffic volume, length and operating environment. Crash-frequency models incorporate these differences when estimating a reference expectation. A frequency-oriented screening objective identifies large crash burdens, whereas an excess-oriented objective identifies sites performing worse than their model-based reference. These questions lead to different rankings and must be specified before evaluating a method [1, 4].

1.2. Network Screening and Methodological Challenges

Crash counts are nonnegative, often overdispersed and unevenly distributed across sites. Road classes may differ in their underlying processes, while adjacent segments can share traffic, environmental conditions and reporting practices. These features complicate both model estimation and evaluation. A statistical procedure that fits aggregate counts adequately may still perform poorly for sparse sites or the high-count tail [5].

The distinction between expected crashes and excess expected crashes is central to this study. An EB expected-frequency score estimates the crash burden associated with a site's posterior risk. An excess score subtracts the model-based reference expectation. Neither score, by itself, estimates the causal reduction achievable through a particular treatment. That calculation requires diagnosis, an appropriate countermeasure and separate effectiveness evidence [4].

1.3. Rationale for Combining EB and Machine Learning

Machine-learning models can represent nonlinear exposure relationships and interactions without requiring every interaction to be specified in advance. Their flexibility is potentially useful when a simple regression mean is misspecified. However, a flexible point prediction is not sufficient to define an EB estimator. The integration must also specify a count likelihood, a distribution for unexplained site variation and a rule for updating that distribution with crash history [6, 7].

Existing research already includes conditional generative adversarial networks within EB estimation and spatial count models with Bayesian additive regression trees. The present study consequently makes no claim to have invented machine-learning-assisted EB. Its purpose is to assess a simpler, auditable integration under a validation design that separates model fitting, dispersion estimation, site-history updating and future evaluation [6, 7].

1.4. Research Gap

Screening methods should be assessed for the decisions they support, rather than judged solely by average prediction error. Earlier evaluation research developed consistency and ranking criteria for hotspot identification. These studies provide a basis for evaluating the selected set of sites as well as the predicted counts [8, 9].

The specific question addressed here is whether a flexible prior mean adds operational value after a conventional EB update has already incorporated recent crash history. The study examines this question at fixed site-investigation budgets, reports calibration and predictive uncertainty, and quantifies a paired difference using corridor-level resampling. This is a focused empirical evaluation, rather than a claim that these dimensions have never been investigated [10].

1.5. Aim and Objectives

The aim is to develop and evaluate an EB and machine-learning framework for estimating crash burden and prioritizing road segments for network-level safety investigation. The objectives are to establish a conventional count-model benchmark, estimate flexible prior means, define their probabilistic EB integration, and compare the resulting forecasts and priority sets on withheld future observations. The analysis also examines dispersion sensitivity and the distinction between site-based and distance-based investigation capacity [3, 4].

1.6. Research Questions

RQ1. How do conventional regression, standalone machine learning and EB-updated models compare when forecasting 2022–2023 reported crash counts on held-out corridors?

RQ2. Does machine-learning-assisted EB capture a larger share of future crashes than conventional EB when investigation is limited to the highest-ranked 5%, 10% or 20% of sites?

RQ3. How much do the selected sets overlap, and how sensitive are the principal comparisons to uncertainty in dispersion and to corridor composition?

RQ4. Which available roadway and exposure variables contribute most to the flexible model's predictions, and what limitations constrain interpretation?

1.7. Contributions and Paper Organization

The study supplies a reproducible secondary analysis with explicit source files, a frozen baseline inventory, a corridor-disjoint evaluation and an EB estimator whose probability model is fully specified. It reports a negative comparative finding when the flexible integration does not improve conventional EB. Sections 2–4 establish the evidence base, data and methods. Sections 5–7 present the results, interpret their scope and identify necessary extensions [3].

2. Literature Review and Conceptual Framework

2.1. Traditional Network Screening Methods

Crash frequency directly represents recorded burden but gives substantial weight to random variation. Crash rate divides events by an exposure measure, which aids comparison but can become unstable where exposure is small. Severity-weighted measures address a different management objective and require credible severity definitions and weights. A method should therefore be selected in relation to the agency's objective and data quality [4].

Comparative hotspot studies demonstrate that methods can produce different sets of priority locations even when they operate on the same records. Evaluation must account for those differences instead of assuming that a high correlation between complete rankings guarantees agreement among the few sites an agency can investigate. Selected-site performance is particularly relevant when capacity is limited [8, 9].

2.2. Safety Performance Functions and EB Estimation

Poisson regression provides a natural starting point for count outcomes, but equality of the conditional mean and variance can be restrictive. Negative-binomial models add a dispersion

parameter that permits variance to exceed the mean. The choice of reference population, exposure specification and dispersion structure consequently affects both predicted crash frequency and the degree of EB shrinkage [2, 5].

In a Poisson–gamma representation, site heterogeneity is described by a latent positive multiplier. Integrating over that multiplier produces a negative-binomial count distribution. Updating the multiplier with crash history produces a posterior expectation between the reference prediction and the history-based estimate. The information supplied by history increases with the expected count in the observation window [2].

2.3. Machine Learning for Crash Prediction and Screening

Random forests average many randomized decision trees and can accommodate nonlinearities and interactions. For count prediction, their outputs remain nonnegative when they average nonnegative training responses. They nevertheless require restrictions on leaf size and independent evaluation to avoid learning unstable local patterns. In this study, a Poisson split criterion is used within the forest rather than treating the model as a classifier [11, 12].

Gradient boosting constructs an additive sequence of learners using a specified loss. A Poisson loss is appropriate for a nonnegative conditional count mean and avoids negative forecasts. The distribution used to define the fitting loss does not automatically become a valid posterior distribution for EB updating. The latter is specified separately in the hierarchical model [12, 13].

2.4. Existing Hybrid Approaches

The Washington freeway study using CGAN-EB evaluated flexible crash-frequency modeling against a conventional NB-EB benchmark. Its existence establishes that replacing a conventional mean model within an EB-oriented framework is an established research direction. The computational and distributional flexibility of generative approaches does not remove the need for transparent validation or comparative screening metrics [6].

Spatial Bayesian additive regression trees provide another approach to combining a flexible mean with count-data inference. This model family explicitly connects flexibility with uncertainty and spatial structure. The present analysis uses a simpler gamma multiplier and corridor-level evaluation, which makes implementation accessible but leaves residual spatial dependence outside the fitted probability model [7].

A subsequent CGAN application examined road-crash risk estimation and hotspot identification, while a 2024 comparison of negative-binomial variants evaluated EB and potential-for-safety-improvement rankings. Together these studies show that the reference distribution, heterogeneity assumptions and ranking objective deserve evaluation alongside algorithm choice [10, 14].

Recent work also considers the consequences of missing traffic volume within fully Bayesian crash-rate estimation. A 2026 preprint develops joint exposure and crash-rate inference, illustrating an active direction for handling uncertainty in traffic data. It is treated here as a preprint, rather than evidence of established performance or peer-reviewed consensus [15].

2.5. Evaluation of Network Screening Methods

Prediction, calibration and screening yield answer different questions. Prediction metrics summarize deviations between forecasts and observed outcomes. Calibration asks whether predicted amounts agree with observed amounts. Screening yield asks how much future crash burden falls within the selected set. A method can improve one dimension and leave another unchanged, especially where selected sites occupy a narrow upper tail of the distribution [8, 9].

Validation must also reflect the intended use. Randomly dividing correlated observations can give an optimistic account of transfer performance. Grouping observations by corridor and withholding future years reduces direct information overlap. Such a design still does not establish transfer to another state or eliminate dependence between nearby corridors [16].

2.6. Synthesis and Conceptual Framework

Table 1 summarizes the literature relevant to the present comparison. Its limitations column identifies boundaries relevant to this study, rather than presenting an exhaustive methodological critique. The framework in Figure 1 separates the role of a flexible prior mean from the role of an independent crash-history update. This separation allows the contribution of each component to be evaluated [6, 7, 10].

Table 1: Literature comparison and scope of the present analysis

Study	Data and units	Methods and integration	Evaluation	Relevant scope
Hauer *et al.* [2]	Road locations; crash history	Poisson–gamma EB	Foundational estimation	Risk estimation; not ML comparison
Cheng and Washington [8]	Hotspot evaluation	Comparative screening criteria	Consistency and ranking tests	Does not evaluate this hybrid
Montella [9]	Roadway hotspot comparison	Alternative screening methods	Method comparison	Different setting and models
Krueger *et al.* [7]	Metropolitan highway counts	Spatial NB with Bayesian trees	Fit and site ranking	More complex spatial inference
Zarei *et al.* [6]	Washington freeways, 2012–2017	CGAN-based EB versus NB-EB	Prediction and screening tests	Different prior distribution and data
Khattak *et al.* [10]	Road-segment crash data	NB variants; EB versus excess	Out-of-sample screening	Different heterogeneity models
Present analysis	Montana segments, 2019–2023	NB/RF/GB means; gamma EB	Corridor holdout; future capture	Baseline exposure; no severity

3. Study Area and Data

3.1. Road Network and Study Setting

The empirical setting is the Montana roadway network represented in the publicly accessible 2019 traffic-count file and 2019–2023 crash extract maintained in the Montana Free Press analysis repository. The repository identifies the records as originating from the Montana Department of Transportation and provides its own matching scripts. This study uses the archived CSV inputs and independently implements the analysis, rather than importing the repository's published crash-rate results [3].

The retained inventory spans 19,117.975 miles across 2,567 corridor identifiers. It contains rural and urban functional groups, including collectors, arterials and interstate categories. This is the network represented by the eligible archived traffic segments. It should not be interpreted as a verified census of every public road in Montana or a contemporary inventory of current roadway conditions [3].

3.2. Analytical Units and Reference Populations

A site is an eligible 2019 traffic segment defined by corridor identifier and start and end reference mileposts. Segment length is taken from SEC_LNT_MI. Reference mileposts define crash assignment, while physical segment length enters the prediction models. These two quantities are kept distinct because reference-marker intervals and physical distance need not be numerically identical [3].

The primary analysis pools sites while including the source traffic functional group as a categorical predictor. This parsimonious benchmark permits differences in the mean among available groups, but uses one residual gamma dispersion per algorithm. It is therefore a locally fitted, limited-covariate SPF benchmark, rather than a direct implementation of every facility-specific equation in the Highway Safety Manual [1, 5].

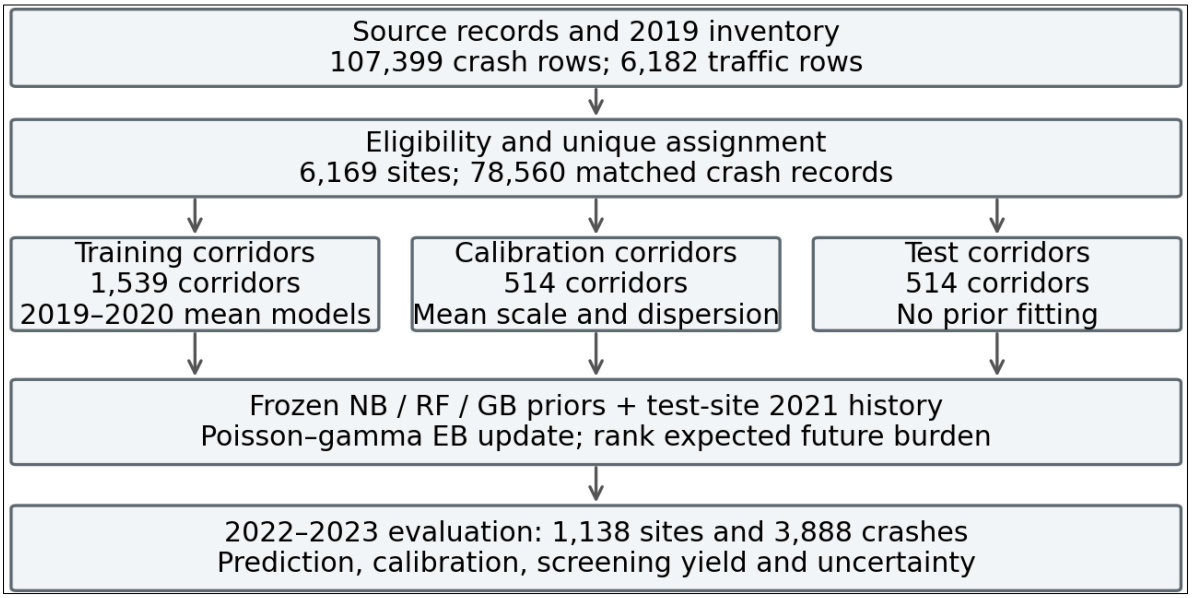


Fig 1: Research workflow and separation of information roles. Schematic drawn for this study

3.3. Data Sources and Observation Periods

The crash extract contains 107,399 rows for 2019–2023. Fields include corridor identifier, reference point, crash year, month, day of week, county and location coordinates. The traffic file contains 6,182 rows with 2019 AADT, physical length, lane count and a traffic functional group. Traffic estimates from later years are deliberately excluded from the predictor set to avoid introducing post-screening information [3].

The historical model-development window is 2019–2020. The EB updating window is 2021, and the screening information cutoff is 31 December 2021. Evaluation covers the next two calendar years, 2022 and 2023. Predictions use the 2019 inventory as a fixed exposure baseline throughout. Accordingly, changes in traffic or roadway configuration after that baseline contribute to forecast error and cannot be separated from changes in underlying crash propensity [3].

3.4. Crash Outcomes

The primary outcome is the number of crash records assigned to each site in a specified period. The extract lacks injury-severity and crash-type fields, so the analysis cannot estimate fatal-and-serious-injury burden, construct defensible severity-weighted scores or evaluate collision-type-specific rankings. A zero count denotes no matched record for the site and period, rather than verified absence of every crash event [3].

3.5. Explanatory Variables and Exposure

All prior-mean models use log AADT, log physical segment length, lane count and the available traffic functional group. Log AADT and log length enter the conventional model with estimated coefficients. Neither is forced to have an elasticity of one. This permits nonproportional relationships while retaining a transparent interpretation of the benchmark [5].

The machine-learning models receive the same predictor information. They do not receive site identifiers, 2021 history counts or future crashes as predictor variables. Excluding history from the prior mean makes the subsequent EB contribution identifiable and avoids counting the same site's history twice through a flexible mean and its posterior update. Available predictor definitions and exclusions appear in Table 2 [3].

Table 2: Variables and sources

Fields	Definition	Source	Analytical use
CORRIDOR / CORR_ID	Crash / traffic corridor identifier	Archived CSVs	Exact matching; not predictive
REF_POINT / CORR_MP / CORR_ENDMP	Reference milepost plus offset	Archived CSVs	Half-open assignment intervals
CRASH_YEAR	Calendar year 2019–2023	Crash CSV	Outcome windows; zero filling
TYC_AADT	2019 vehicles per day	Traffic CSV	Positive; log predictor
SEC_LNT_MI	Physical segment miles	Traffic CSV	Positive; log predictor
NUM_LANES	Recorded lane count	Traffic CSV	Positive numerical predictor
FACTOR_GRP	Traffic functional group	Traffic CSV	Trimmed categorical predictor
Severity / crash type	Not provided	Not available	No severity or type analysis

3.6. Data Linkage and Quality Assurance

Reference points of the form 030+0.926 are parsed as 30.926. Crash assignment requires an exact corridor match and a reference point in the half-open interval from the segment start, inclusive, to its end, exclusive. The rule prevents an event exactly at an adjoining boundary from being assigned to both segments. All remaining overlapping eligible traffic segments are removed before crash matching, and the code asserts that no record receives more than one assignment [3].

Seven traffic rows fail the required validity checks, and six additional segments are removed because of overlap. No eligible rows collapse as exact duplicates under the selected segment and predictor keys. The final inventory therefore contains 6,169 sites. Of the original crash rows, 78,560 are assigned and 28,839 remain unmatched. Unmatched records are excluded from site outcomes, with no assumption that their omission is random [3].

The crash file contains 576 exact duplicate rows across its available fields. These rows are retained because the extract has no unique crash identifier and two distinct events can share the same disclosed attributes. Retention is transparent but does not establish that every repeated row is a separate crash. Agency identifiers would be needed to resolve duplication definitively. Missing coordinates do not cause exclusion because matching uses corridor and reference point [3].

3.7. Descriptive Analysis

Table 3 reports the corridor-disjoint partitions and yearly matched counts. The training set contains 3,920 sites, the calibration set 1,111 sites and the test set 1,138 sites. The partitions differ in exposure and count distributions. These differences are retained as part of the transfer evaluation rather than corrected using future outcomes. Figure 2 presents observed annual totals across all retained sites [3].

Table 3: Corridor partitions and matched crash counts

Partition	Sites	Corridors	2019–20 crashes	2021 crashes	2022–23 crashes
train	3920	1539	23714	12520	23065
calibration	1111	514	3737	1963	3558
test	1138	514	4025	2090	3888

4. Methodology

4.1. Overall Research Design

The experiment isolates three information sources: roadway and exposure features, the separate crash history used for updating, and future observations used only for evaluation. Whole corridor identifiers are assigned to training, calibration and test partitions with fixed random seeds. The nominal proportions are 60%, 20% and 20% of corridors, not of sites. Unequal numbers of sites per corridor produce different site proportions [16].

Within training corridors, three-fold GroupKFold validation selects each machine-learning configuration using mean Poisson deviance for 2019–2020 aggregate counts. Mean calibration and dispersion estimation then use the separate calibration corridors. All final metrics and ranking comparisons use the test corridors and their 2022–2023 outcomes. No model is selected using those future test crashes [16, 17].

4.2. Conventional Statistical Benchmark

Let C_i denote a site's aggregate crash count in 2019–2020 and m_i its predicted annual mean. The benchmark uses a negative-binomial log-link model with $\mathrm{Var}(C_i) = M_i + a_{train} M_i^2$, where $M_i = 2m_i$. The fitted predictor is $\log M_i = \log 2 + b_0 + b_1 \log \mathrm{AADT}_i + b_2 \log L_i + b_3 \mathrm{lanes}_i$ + class effects. Coefficients and training dispersion are estimated by maximum likelihood [5].

The temporal offset log 2 converts an annual mean to the two-year fitting count. Segment length and AADT remain estimated predictors rather than offsets. The recreational traffic group is the alphabetical reference category in the training vocabulary. Coefficients describe conditional predictive associations within the available inventory. Their standard errors do not account for all spatial dependence or support causal claims [3, 5].

4.3. Conventional EB Estimation

For an annual reference mean m_i, let the latent multiplier U_i follow Gamma(r,r), using shape and rate, with r = 1/alpha. Thus $E(U_i) = 1$ and $\mathrm{Var}(U_i) =$ alpha. Conditional on U_i, a period with baseline expected count M has a Poisson count with mean U_i M. The marginal period count is negative binomial, with variance M + alpha M^2 [2].

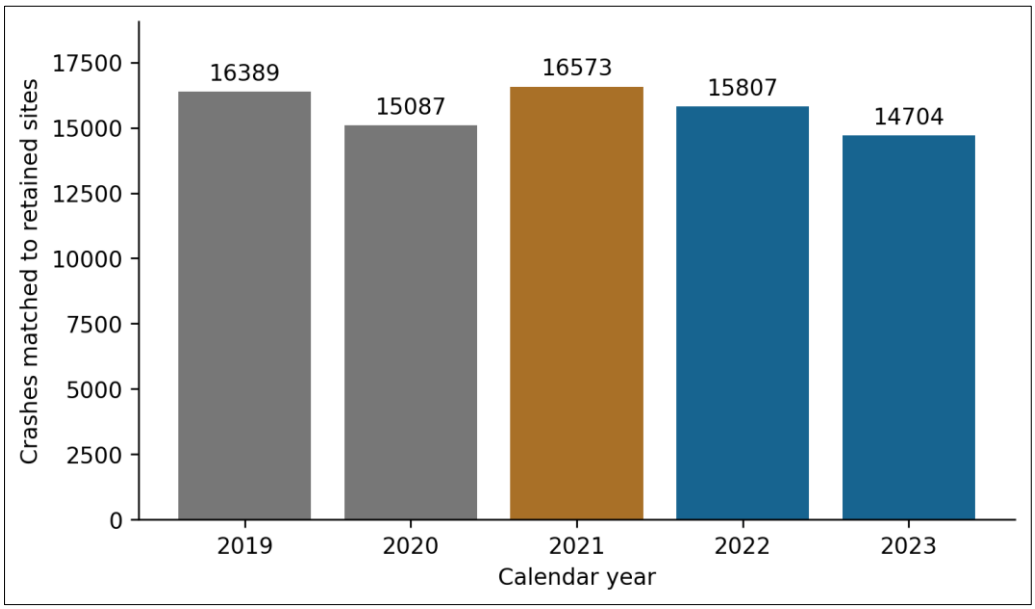


Fig 2: Observed annual crashes on the retained network.

For H_i crashes observed during the 2021 history window, the posterior multiplier is Gamma($r + H_i$, $r + m_i$). The posterior expected count in that history window is $m_i(r + H_i)/(r + m_i)$. Equivalently, it is $w_i m_i + (1 - w_i)H_i$, where $w_i = 1/(1 + \text{alpha } m_i)$. The weight therefore follows from the model's heterogeneity distribution and expected history exposure [2].

For the two-year future period, the forecast is $P_i = 2m_i(r + H_i)/(r + m_i)$. This assumes the same latent site multiplier persists from the history to the forecast period and that the baseline annual mean remains applicable. More generally, with time-varying available exposure, m_i in the denominator is replaced by the sum of history-period mean predictions and $2m_i$ by the future sum. Such exposure changes are not estimated here [2, 3].

4.4. Machine-Learning Model Development

The random forest contains 200 trees and uses the Poisson splitting criterion. Candidate minimum leaf sizes are 10 and 30, and candidate feature fractions are 0.7 and 1.0. All other unreported settings retain the documented defaults of the recorded software version. The selected forest uses minimum leaf size 10 and all available features at each split [11, 12].

Histogram gradient boosting uses Poisson loss, learning rate 0.05, 150 iterations, minimum leaf size 30 and disabled early stopping. Candidate maximum leaf counts are 7 and 15, with L2 regularization 1 or 10. Training-corridor validation selects seven leaves and regularization 10. The target is the annualized 2019–2020 count, while validation predictions and outcomes are multiplied by two to evaluate the fitting-period aggregate [12, 13].

Categorical indicators use the training vocabulary only, with one reference category omitted. Continuous variables are not standardized because tree models do not require that transformation and the regression uses interpretable log predictors. Any category absent from the training vocabulary would map to the reference representation. This conservative implementation should be replaced by an explicit unknown-category policy before operational use [3, 17].

4.5. Integration of Machine Learning and EB

For each model, a single multiplicative calibration factor is estimated on calibration corridors using the ratio of observed to predicted 2019–2020 totals. The resulting annual means define $M_i = 2m_i$ on those corridors. A model-specific alpha is then estimated by maximizing the negative-binomial likelihood of the calibration aggregate counts, with mean M_i and shape 1/alpha. The optimization operates on log alpha within [−9,4]. No test-site count enters these parameter estimates [3].

The calibrated NB, forest and boosting means each enter the same Poisson–gamma hierarchy described above. This defines three EB methods, NB-EB, RF-EB and GB-EB, together with their standalone prior-mean forecasts. A history-only comparator predicts twice the observed 2021 count. By holding the updating rule fixed, the comparison asks whether the flexible prior supplies useful information beyond the conventional mean and recent history [3, 6].

This integration is semi-parametric only in its mean component. Its gamma heterogeneity distribution remains parametric, common across sites within an algorithm, and stationary across time. Tree-to-tree variation is not used as a substitute for unexplained site heterogeneity, and the Poisson fitting loss is not interpreted as evidence that site counts lack overdispersion [5, 7].

4.6. Screening Measures and Prioritization Rules

The primary ranking score is expected 2022–2023 crash burden. Sites are sorted in descending score, with lexicographic site identifier as the deterministic tie-breaker. At budget b, the selected size is ceil(bn), where n is the number of test sites. Results emphasize 5%, 10% and 20% budgets, while the downloadable table includes additional capacities [3, 8].

Excess expected counts are also supplied in the analytic panel as $P_i - 2m_i$ for NB-EB and GB-EB. Negative values are retained rather than truncated. These columns are auxiliary screening indicators and are not evaluated as if they optimize the primary burden-capture objective. No treatment-specific crash savings are inferred from them [4].

4.7. Validation and Prevention of Data Leakage

A test site's 2021 history legitimately enters its EB update because it is available at the screening cutoff. Its 2019–2020 counts do not enter model fitting or calibration because its

entire corridor is withheld. Its 2022–2023 counts remain unavailable until evaluation. The design thus combines transfer of a prior to unseen corridors with retrospective forecasting at sites for which recent crash history exists [16].

The use of a retrospectively assembled archive requires a separate qualification. The information cutoff refers to the event years and baseline predictor years represented in the analysis. The repository does not establish exactly when every record became available to an agency, or whether later administrative revisions altered historical fields. This is retrospective validation, not a prospective deployment trial [3].

4.8. Evaluation Metrics

Prediction metrics are MAE, RMSE and mean Poisson deviance on two-year site counts. Calibration is summarized by the observed-to-predicted total ratio and observed versus predicted means within prediction deciles. A ratio above one indicates aggregate underprediction. Spearman correlation describes association between forecast ranks and future observed-count ranks, with future counts acknowledged as noisy outcomes [3, 8].

Screening yield is 100 times the sum of future crashes at selected sites divided by all future test crashes. Selection overlap is the fraction of selected sites shared with NB-EB at the same budget. These measures are calculated directly from the held-out outcomes. Precision and recall are omitted because the dataset supplies no independent binary label of latent hotspot status [8, 9].

The posterior predictive distribution for a future count is negative binomial with shape $r + H_i$ and success probability $(r + m_i)/(r + m_i + 2m_i)$. Central 95% count intervals are obtained from its discrete quantiles. Their coverage and mean width evaluate conditional predictive uncertainty. They do not include uncertainty in fitted model coefficients, calibration factors, alpha or future exposure [2, 3].

4.9. Model Interpretation

Boosting interpretation uses permutation importance on calibration corridors, repeating each permutation ten times and measuring the increase in annual-count Poisson deviance. The calculation measures predictive reliance in the fitted model. It does not identify the causal effect of changing traffic, road length or lane count, and correlated predictors can alter the distribution of importance among variables [17].

4.10. Sensitivity and Uncertainty Analyses

The principal sensitivity analysis multiplies each estimated alpha by 0.5 and 2 while retaining the calibrated mean and history counts. This tests how shrinkage assumptions affect prediction and 10% screening yield. It is an exploratory diagnostic, not a procedure for selecting alpha after viewing test outcomes. Functional-group errors and calibration ratios are also supplied as supplementary outputs [3, 10].

A paired bootstrap resamples the 514 test corridors with replacement 1,000 times. Each sampled corridor contributes all of its sites, each method is reranked within the resample and the top 10% selected anew. Percentile intervals summarize the GB-EB minus NB-EB differences in crash capture and MAE. Models remain fixed, so the intervals reflect evaluation-sample composition rather than the full model-development process [3, 16].

Segmentation alternatives, multiple independently reserved temporal origins, exposure imputation and full refitting under a spatial bootstrap are not implemented. These boundaries are retained explicitly instead of presenting the available sensitivity analysis as a complete examination of ranking stability. The supplied code and panel permit those extensions when additional source information is available [3].

4.11. Reproducibility and Implementation

The supplementary package contains the analysis script, source snapshots, analytic site panel, complete metric tables, plotting data, fitted coefficients, tuning scores, source hashes and software versions. The random seed is 260516, with 260517 used for the second partition. The implementation uses Python 3.12.14, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0 and statsmodels 0.15.0. Reported results were computed directly rather than transcribed from earlier papers [3, 17].

5. Results

5.1. Final Dataset and Descriptive Findings

The eligibility and linkage procedures retain 6,169 traffic segments and 78,560 crash rows, corresponding to 73.15% of the original crash extract. The unmatched 26.85% is substantial, and the results are therefore interpreted for the retained network only. Training, calibration and test corridor counts are 1,539, 514 and 514. The test sites span 3,262.494 miles and contain 3,888 evaluation crashes [3].

Across all retained sites, annual matched crashes total 16,389 in 2019, 15,087 in 2020, 16,573 in 2021, 15,807 in 2022 and 14,704 in 2023. These are descriptive totals from a fixed segment inventory. They do not establish exposure-adjusted changes or the effect of any safety intervention. In the test partition, 53.78% of sites have no matched crash in the 2021 history window [3].

5.2. Statistical Model and EB Estimation Results

The NB maximum-likelihood optimizer converges. The log AADT coefficient is 0.835, with model-based standard error 0.019, while log length is 0.678, with standard error 0.017. These subproportional fitted associations explain why directly dividing crash frequency by vehicle-miles would impose a stronger exposure assumption than the estimated benchmark. The lane coefficient is 0.056, with standard error 0.029 [3].

The training NB2 dispersion is 0.624, while independently calibrated alpha is 0.952 for NB, 0.974 for RF and 0.938 for GB. The values arise from different samples and roles: training dispersion accompanies coefficient fitting, while calibration dispersion defines the EB prior used in forecasting. Mean calibration multipliers are approximately 0.980, 0.975 and 0.984. Median test-site prior weights are approximately 0.555 for all three methods [3].

Table 4: Calibration and gamma prior parameters

Prior	Mean multiplier	Calibrated alpha	Median prior weight
NB	0.980	0.952	0.555
RF	0.975	0.974	0.555
GB	0.984	0.938	0.556

5.3. Machine-Learning Prediction Results

Among standalone means, GB has the lowest RMSE, 4.419, compared with 4.521 for NB and 4.484 for RF. GB's MAE is 2.309, compared with 2.329 for NB. The RMSE reduction relative to NB is approximately 2.25%. These point comparisons establish a modest difference in the evaluated sample; no full-refit confidence interval for this standalone difference is estimated [3].

Standalone forecasts underpredict the evaluation aggregate. Observed-to-predicted ratios are 1.051 for NB, 1.076 for RF and 1.058 for GB. Calibration on earlier calibration corridors consequently does not guarantee agreement for later test-corridor outcomes. The calibration and error plots in Figure 3 should be interpreted alongside the aggregate ratios in Table 5 [3].

Table 5: Future test prediction and calibration performance

Model	MAE	RMSE	Deviance	Observed / predicted	95% coverage
NB	2.329	4.521	2.885	1.051	
NB-EB	1.875	3.650	2.044	0.946	0.974
RF	2.312	4.484	2.943	1.076	
RF-EB	1.882	3.669	2.071	0.955	0.966
GB	2.309	4.419	2.863	1.058	
GB-EB	1.887	3.661	2.046	0.944	0.972
History	2.139	4.207	16.595	0.930	

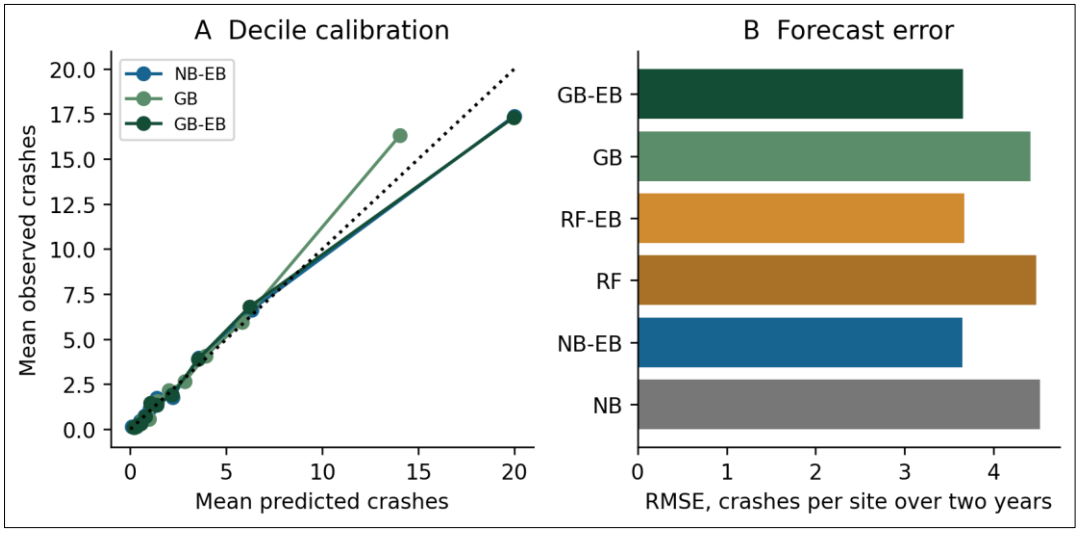


Fig 3: Test calibration and forecast error. Panel A uses test-site prediction deciles and two-year counts.

5.4. Hybrid Framework Performance

EB updating substantially improves each prior-mean model's point errors. NB-EB reduces MAE from 2.329 to 1.875 and RMSE from 4.521 to 3.650. GB-EB reduces the corresponding standalone GB values from 2.309 to 1.887 and from 4.419 to 3.661. However, GB-EB does not improve on NB-EB. RF-EB has MAE 1.882 and RMSE 3.669 [3].

The paired corridor-bootstrap interval for the GB-EB minus NB-EB MAE difference is −0.0047 to 0.0292 crashes per site. This interval includes zero. NB-EB, RF-EB and GB-EB have predictive interval coverage of 97.36%, 96.57% and 97.19%, respectively. Their mean interval widths are 8.03, 7.94 and 8.08 crashes. Coverage above 95% is compatible with conservative discrete intervals and does not establish accurate tail uncertainty at every site [3].

EB forecasts overpredict the future aggregate, with observed-to-predicted ratios of 0.946 for NB-EB, 0.955 for RF-EB and 0.944 for GB-EB. The direction differs from that of the standalone models. Recent history therefore adds useful site information while also carrying a temporal level that is high relative to the subsequent period. The analysis does not identify the causes of that difference [3].

5.5. Network Screening and Prioritization Results

At the 10% budget, 114 sites are selected. NB-EB captures 1,981 of 3,888 future crashes, or 50.95%. RF-EB captures 1,982, or 50.98%, and GB-EB captures 1,973, or 50.75%. History-only ranking captures 1,962 crashes, or 50.46%. The flexible EB methods therefore produce very similar selected-site yields to conventional EB, rather than a substantial operational gain [3].

GB-EB differs from NB-EB by −0.206 percentage points in 10% crash capture. The paired bootstrap interval is −0.679 to 0.853 percentage points. RF-EB's one-crash point advantage is descriptive and is not accompanied by a separate inferential comparison. Table 6 reports additional budgets, and Figure 4 shows how the yield changes with investigation capacity [3].

At the 10% budget, both RF-EB and GB-EB share 113 of their 114 selected sites with NB-EB, an overlap of 99.12%. NB-EB's selected set spans 306.921 miles, while GB-EB's spans 305.535 miles. A 10% site budget therefore is not equivalent to a 10% mileage budget. The near-complete selected-set overlap helps explain why the observed screening yields are close [3].

Table 6: Future crash capture at fixed site investigation budgets

Model	Budget %	Sites	Crashes captured	Capture %	NB-EB overlap %
NB	5	57	1229	31.610	68.421
NB	10	114	1808	46.502	63.158
NB	20	228	2555	65.715	71.930
NB-EB	5	57	1385	35.622	100.000
NB-EB	10	114	1981	50.952	100.000
NB-EB	20	228	2736	70.370	100.000
RF	5	57	1226	31.533	66.667
RF	10	114	1831	47.094	65.789
RF	20	228	2534	65.175	69.298
RF-EB	5	57	1388	35.700	98.246
RF-EB	10	114	1982	50.977	99.123
RF-EB	20	228	2736	70.370	96.491
GB	5	57	1244	31.996	66.667
GB	10	114	1858	47.788	67.544
GB	20	228	2537	65.252	70.175
GB-EB	5	57	1398	35.957	98.246
GB-EB	10	114	1973	50.746	99.123
GB-EB	20	228	2747	70.653	97.807
History	5	57	1373	35.314	92.982
History	10	114	1962	50.463	93.860
History	20	228	2676	68.827	90.351

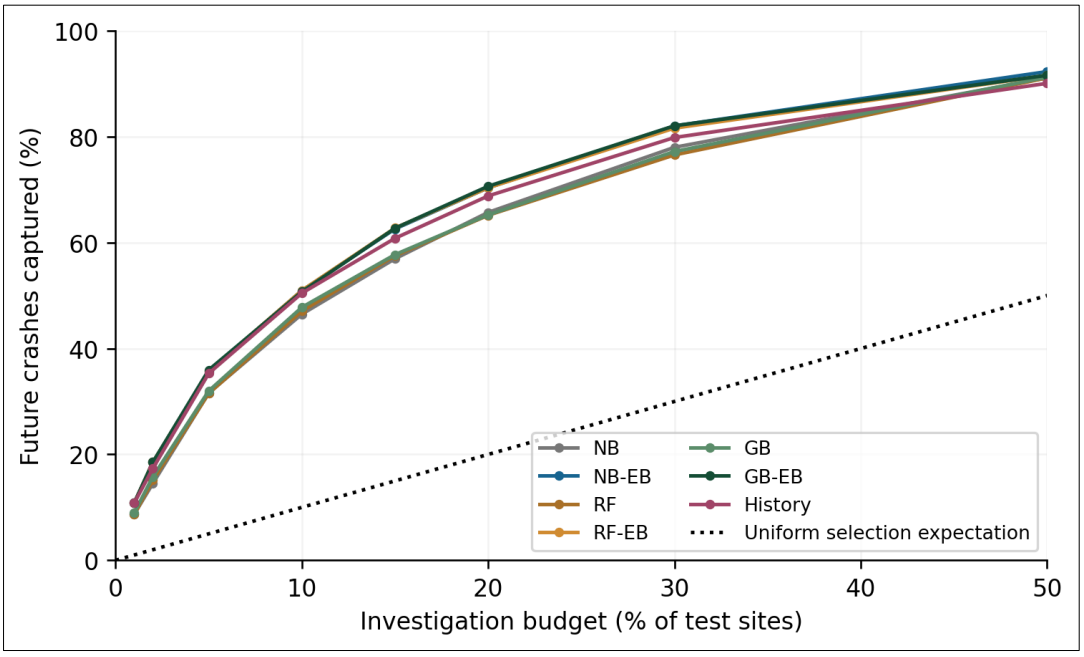


Fig 4: Future crash capture across site investigation budgets.

Figure 5 compares site-level NB-EB and GB-EB predictions. Agreement is strong in the upper-ranked selected set but not exact in predicted counts. The plot contains only test sites and does not use coordinates to imply a geographic pattern. Detailed priority locations can be identified through the corridor and milepost fields in the supplied panel, subject to source verification before field investigation [3].

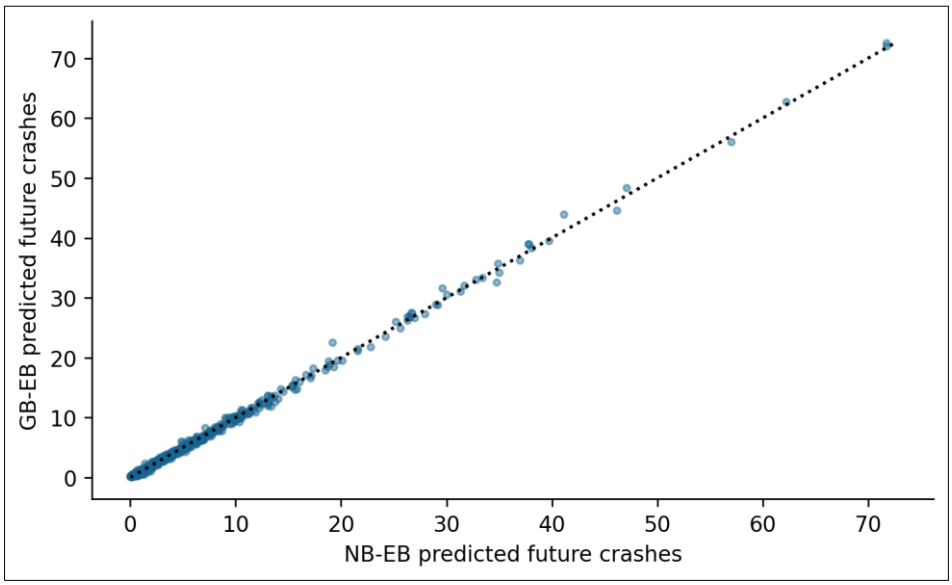


Fig 5: Conventional and boosting-assisted EB forecasts.

5.6. Model Interpretation

Permutation of log AADT increases annual-count Poisson deviance by 4.681 on average, while permutation of log length increases it by 3.788. Lane count contributes 0.023 under the same diagnostic. Individual class indicators contribute little additional importance in this fitted model, with some negative estimates. Such small values can arise from sampling variation, redundant predictors and model structure [3].

The model therefore relies mainly on traffic and physical segment extent within the observed feature set. This result does not mean that engineering features absent from the extract are unimportant. It also does not establish the consequence of changing traffic or segment boundaries. In particular, length is partly an analytical-unit property, so its importance reinforces the need to report segmentation explicitly [3].

5.7. Sensitivity and Uncertainty Results

Halving alpha improves point MAE for all three EB methods and changes their ranking at the 10% screening budget. GB-EB capture rises to 51.47%, compared with 51.05% for NB-EB. Doubling alpha yields 50.95% capture for both methods. These exploratory comparisons show that the apparent advantage of a prior-mean algorithm can depend on shrinkage specification [3].

Table 7: Sensitivity to gamma dispersion

Model	Alpha multiplier	MAE	10% capture %
NB-EB	0.500	1.843	51.055
RF-EB	0.500	1.858	51.235
GB-EB	0.500	1.860	51.466
NB-EB	1.000	1.875	50.952
RF-EB	1.000	1.882	50.977
GB-EB	1.000	1.887	50.746
NB-EB	2.000	1.943	50.952
RF-EB	2.000	1.944	50.720
GB-EB	2.000	1.951	50.952

These sensitivity findings do not justify replacing the pre-test calibrated alpha with a value chosen from evaluation outcomes. They instead identify dispersion modeling as a priority for subsequent validation. The supplementary functional-group results also reveal uneven sample sizes, so small-group performance is descriptive. No uniform subgroup superiority or external-network transfer advantage is established [3].

6. Discussion

6.1. Principal Findings

RQ1 is answered by the distinction between standalone and updated forecasts. Flexible boosting provides a modest standalone RMSE improvement, while NB-EB gives the smallest evaluated MAE and RMSE among the EB methods. RQ2 receives a negative comparative answer: GB-EB does not improve future crash capture over NB-EB at the primary 10% investigation budget, and the paired interval includes no difference [3].

For RQ3, selected-site overlap is extremely high, while dispersion sensitivity shows that small ranking differences can change direction. The bootstrap intervals describe uncertainty associated with corridor composition, rather than complete model uncertainty. RQ4 identifies traffic and segment length as the main available predictive inputs, with interpretation limited by the small feature set and the absence of causal identification [3].

6.2. Interpretation of Method Differences

The flexible prior and the EB update contribute different information. Boosting can modify relationships between exposure and the prior expected count. The history update then adjusts each site using its recent observed crashes. Where recent history is highly informative for the upper-ranked sites, alternative prior means can converge to nearly the same operational selection. That interpretation is consistent with the 99.12% overlap observed here [3].

The results also demonstrate that a reduction in global forecast error does not guarantee a better selected set. RMSE is sensitive to larger prediction errors throughout the network, whereas top-budget capture depends on a restricted upper-ranked set. Conventional ranking-evaluation research motivates examining both dimensions; this dataset provides a concrete example of their different responses to model flexibility [8, 9].

6.3. Comparison with Previous Research

Earlier CGAN-EB work provides evidence that flexible modeling can improve crash-frequency modeling and screening under its evaluated conditions. The present result does not contradict the existence of those gains. The tested integration uses a gamma heterogeneity model, a narrower covariate set, a different network and a corridor-disjoint future evaluation. These differences prevent a universal conclusion about the value of generative or other flexible models [6, 14].

The dispersion sensitivity is also consistent with the broader importance of modeling heterogeneity in EB analysis. Research on alternative negative-binomial formulations and spatial Bayesian trees considers aspects of variation that a shared gamma multiplier cannot fully describe. The present findings support evaluating those assumptions directly rather than treating algorithm flexibility as sufficient evidence of better screening [7, 10].

6.4. Methodological Contribution

The contribution is a transparent comparison in which prior estimation, calibration, updating and evaluation have separate data roles. The approach exposes a case where adding a flexible prior yields little additional screening value. It also provides a reproducible distinction between burden-based rankings and auxiliary excess scores. Its originality lies in the executed analysis and documented comparison, rather than in claiming a new EB identity [3].

6.5. Implications for Road Safety Management

An agency with data resembling this extract would have limited empirical reason to replace the tested conventional EB benchmark solely on the basis of the flexible model's standalone prediction improvement. It could retain the simpler model while investing in better exposure updates, crash identifiers and severity fields. This is a network-specific interpretation of the evaluated alternatives, not a recommendation against machine learning in other settings [3].

Operational implementation should begin by specifying the objective and investigation unit. If the agency's priority is severe-injury burden, these total-record results are insufficient. If investigation cost depends on distance or junction complexity, site-count budgets also require modification. Screening should then lead to engineering diagnosis and countermeasure assessment, with uncertainty communicated alongside the ranked list [4].

6.6. Transferability

The held-out corridors assess transfer within the archived Montana inventory. They do not establish performance for another state, another country's reporting system, intersections as separate analytical units or current road conditions. Application elsewhere would require a compatible segment inventory, a local exposure reference and independent calibration of the residual distribution [3].

Exposure uncertainty is a particularly important transfer constraint. The fixed 2019 AADT baseline is operationally reproducible but cannot represent all later changes. A future extension could model annual exposure and its uncertainty, potentially using a hierarchical approach. Recent preprint work provides a relevant methodological direction, although it has not been validated as part of the present experiment [15].

6.7. Limitations

The first limitation is data provenance and completeness. The analysis uses a public repository copy of agency-origin records, not an independently reconciled agency release. Unmatched crashes constitute 26.85% of the extract, exact duplicate rows cannot be conclusively resolved, and reporting completeness is unknown. The study therefore estimates the burden represented in the matched records rather than all true road crashes [3].

The second limitation concerns exposure, segmentation and reference groups. Annual traffic changes, roadway reconstruction, detailed alignment, access density, lighting and roadside conditions are unavailable or unused. Segment boundaries are inherited from the 2019 inventory. All functional groups share one residual dispersion within each algorithm, and the recreational reference category may not be an ideal engineering reference population [3].

The third limitation is inferential scope. Corridor grouping reduces direct overlap but does not fully model spatial correlation or temporal changes. The fitted gamma distribution assumes a persistent multiplier, and posterior predictive intervals condition on fitted parameters. Bootstrap comparisons do not refit the models. Only one primary temporal forecast origin and one network are evaluated [3, 16].

The fourth limitation is the outcome and decision target. Severity, exposure-normalized current risk and treatment effectiveness are not measured. Future crash counts are noisy realizations, not error-free labels of latent hotspots. The analysis cannot establish prevented crashes, cost effectiveness, or superiority of a method for selecting a particular countermeasure [3, 4].

7. Conclusions

This study evaluated conventional and machine-learning-assisted EB screening using 6,169 Montana traffic segments and a future test period containing 3,888 crashes across 1,138 sites. A Poisson–gamma hierarchy connected calibrated prior means with a separate 2021 history update. Entire corridors were withheld from prior fitting and calibration, and all final forecasting comparisons used 2022–2023 outcomes [3].

Boosting improved standalone RMSE from 4.521 to 4.419, but conventional NB-EB achieved lower MAE than GB-EB, 1.875 versus 1.887. At a 10% site budget, NB-EB captured 50.95% of future crashes and GB-EB 50.75%. The estimated GB-EB difference of −0.21 percentage points had a corridor-bootstrap interval spanning −0.68 to 0.85. The tested hybrid therefore offers no demonstrated screening advantage over conventional EB on this network [3].

The practical contribution is a reproducible evaluation that separates prediction quality from investigation yield and makes the probabilistic integration explicit. Future research should prioritize updated exposure, verified crash identifiers, severity-specific outcomes, alternative dispersion structures, multiple reserved time origins and validation on an external network. Any claim of actual crash reduction would require a separately designed intervention evaluation [3, 4].

References

1. American Association of State Highway and Transportation Officials. Highway Safety Manual. First edition. AASHTO, Washington, DC; 2010. Available from: https://www.highwaysafetymanual.org/pages/default.aspx

2. Hauer E, Harwood DW, Council FM, Griffith MS. Estimating Safety by the Empirical Bayes Method: A Tutorial. Transportation Research Record: Journal of the Transportation Research Board. 2002;1784(1):126-131. doi:10.3141/1784-16
3. Montana Free Press. MDT segment level data analysis. Archived Montana Department of Transportation crash and traffic CSV inputs. Repository snapshot a3de6de187f6f30dc7c0681d65c1fb61c1e887b1; 2025. Available from: https://github.com/mtfreepress/mdt-segment-data-analysis [cited 16 October 2026]
4. Srinivasan R, Gross F, Lan B, Bahar G. Reliability of Safety Management Methods: Network Screening. Report FHWA-SA-16-037. Federal Highway Administration, Washington, DC; 2016. Available from: https://highways.dot.gov/sites/fhwa.dot.gov/files/2022-09/fhwasa16037.pdf
5. Lord D, Mannering F. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice. 2010;44(5):291-305. doi:10.1016/j.tra.2010.02.001
6. Zarei M, Hellinga B, Izadpanah P. CGAN-EB: A non-parametric empirical Bayes method for crash frequency modeling using conditional generative adversarial networks as safety performance functions. International Journal of Transportation Science and Technology. 2023;12(3):753-764. doi:10.1016/j.ijtst.2022.06.006
7. Krueger R, Bansal P, Buddhavarapu P. A new spatial count data model with Bayesian additive regression trees for accident hot spot identification. Accident Analysis & Prevention. 2020;144:105623. doi:10.1016/j.aap.2020.105623
8. Cheng W, Washington S. New Criteria for Evaluating Methods of Identifying Hot Spots. Transportation Research Record: Journal of the Transportation Research Board. 2008;2083(1):76-85. doi:10.3141/2083-09
9. Montella A. A comparative analysis of hotspot identification methods. Accident Analysis & Prevention. 2010;42(2):571-581. doi:10.1016/j.aap.2009.09.025
10. Khattak MW, De Backer H, De Winne P, Brijs T, Pirdavani A. Comparative Evaluation of Crash Hotspot Identification Methods: Empirical Bayes vs. Potential for Safety Improvement Using Variants of Negative Binomial Models. Sustainability. 2024;16(4):1537. doi:10.3390/su16041537
11. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32. doi:10.1023/a:1010933404324
12. scikit-learn developers (2026). Ensemble estimator API documentation. RandomForestRegressor and HistGradientBoostingRegressor. Accessed 16 October 2026. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html; https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html
13. Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001;29(5):1189-1232. doi:10.1214/aos/1013203451
14. Zarei M, Hellinga B, Izadpanah P. Application of Conditional Deep Generative Networks (CGAN) in empirical bayes estimation of road crash risk and identifying crash hotspots. International Journal of Transportation Science and Technology. 2024;13:258-269. doi:10.1016/j.ijtst.2023.02.005
15. Skaug L. Beyond Empirical Bayes: A Hierarchical Bayesian Approach to Crash Rate Estimation with Missing Traffic Volume. arXiv:2605.27889v2. Preprint, revised. 2026 Jun 1. doi:10.48550/arXiv.2605.27889
16. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, *et al*. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913-929. doi:10.1111/ecog.02881
17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825-2830. Available from: https://www.jmlr.org/papers/v12/pedregosa11a.html

How to Cite This Article

Afolabi AO, Omoifo-Irefo MO. Applying empirical bayes methods and machine learning to Network-Level road safety screening. Int J Revolut Civ Eng. 2026;2(6):1-12. doi:10.54660/IJRCE.2026.2.6.01-12.

Creative Commons (CC) License

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.