

When Do Logistic Regression and Generalized Maximum Entropy Agree? Evidence from Simulation Studies

Maharaj Singh

College of Nursing, Marquette University, Milwaukee, WI 53233, USA; maharaj.singh@marquette.edu

How To Cite: Singh, M. When Do Logistic Regression and Generalized Maximum Entropy Agree? Evidence from Simulation Studies. *Journal of Modern Applied Statistical Methods* 2026, 25(2), 12. <https://doi.org/10.53941/jmasm.2026.100012>

Received: 31 December 2025

Revised: 6 May 2026

Accepted: 28 July 2026

Published: 7 August 2026

Abstract: Logistic regression is widely used for binary outcomes, yet generalized maximum entropy (GME) estimators are often recommended when data are limited or ill-conditioned. Applied guidance on when GME and logistic regression provide practically similar predictive performance—particularly probability calibration—remains limited. We conducted a Monte Carlo study comparing SAS PROC LOGISTIC (maximum likelihood) with SAS PROC ENTROPY (GME-D; GMED) across four baseline scenarios: (S1) large well-specified samples ($n = 5000$), (S2) small samples with high collinearity ($n = 100$; $\text{corr}(x_1, x_2) \approx 0.95$), (S3) rare events with a quasi-separation tendency ($n = 1000$; $\beta_0 = -3.50$; 1% of observations with x_1 shifted by +12), and (S4) heavy-tailed contamination/outliers ($n = 1000$; 5% contaminated with $t(1)$ noise). Each condition was replicated $R = 200$ times with a 70/30 train–test split. Performance was evaluated on test data using discrimination (AUC), proper scoring rules (Brier score, log loss), and calibration via logistic recalibration (intercept and slope). Across S1–S2, AUC and scoring rules were nearly identical between methods, but PROC ENTROPY produced modestly larger calibration slopes. In S3, AUC remained similar ($\Delta\text{AUC} \approx 0.0014$), while calibration diverged substantially ($\Delta \text{intercept} \approx 0.456$; $\Delta \text{slope} \approx 0.278$). In S4, PROC LOGISTIC showed better discrimination ($\Delta\text{AUC} \approx -0.0055$) and slightly better scoring rules, while PROC ENTROPY again produced larger calibration slopes. Sensitivity analyses varying event prevalence, ESUPPORTS width, and sample size confirmed that calibration differences are most pronounced under rare events and depend on entropy support specifications. Routine calibration reporting is essential when comparing binary classifiers and PROC ENTROPY users should assess sensitivity to ESUPPORTS choices when probability estimation is the goal.

Keywords: generalized maximum entropy; PROC ENTROPY; logistic regression; PROC LOGISTIC; calibration; discrimination; rare events; outliers; simulation

1. Introduction

Binary outcomes are ubiquitous in health, behavioral, and social science research (e.g., disease presence/absence, readmission, treatment response, adverse events). Logistic regression remains the dominant approach because it links predictors to event probabilities in an interpretable way and is efficiently estimated via maximum likelihood [1,2]. For prediction and risk stratification, model performance is commonly summarized by discrimination (AUC), overall probabilistic accuracy via proper scoring rules (Brier score and log loss), and calibration (agreement between predicted and observed risks) [3–6].

Maximum-likelihood logistic regression can be challenged in several applied settings. Small samples may yield unstable estimates and finite-sample bias [7]; rare events and low events-per-predictor settings can distort probabilities [8]; complete or quasi-complete separation can prevent finite estimates [9,10]; and collinearity and



influential observations can destabilize coefficients and predictions [11,12]. These limitations motivate alternatives such as penalization, bias-reduction, and Bayesian approaches.

Although generalized maximum entropy has been proposed as an attractive alternative when information is limited or data are ill-conditioned [13,14], relatively little guidance is available regarding when GME produces practically similar predictive results to conventional logistic regression. To our knowledge, few studies have systematically compared these methods using modern out-of-sample performance measures including discrimination, proper scoring rules, and probability calibration—across multiple realistic data scenarios. Consequently, practitioners have limited evidence regarding when the two approaches may be used interchangeably for prediction and risk estimation.

Generalized maximum entropy (GME) provides an information-theoretic alternative to likelihood-based estimation. Under the maximum entropy principle, estimation selects the least-informative coefficient representation consistent with the observed data constraints; in regression settings, parameters are represented on finite supports with associated probability weights [13]. In SAS, PROC ENTROPY implements GME methods for regression models and can be particularly useful when conventional estimation is challenged by limited information or ill-conditioned designs. Throughout this paper, “agreement” refers to practical similarity in predictive performance including discrimination, proper scoring rules, and calibration rather than exact equality of parameter estimates or identical numerical performance on individual metrics.

For applied analysts, the central question is not only whether GME converges, but whether it yields practically equivalent out-of-sample predictions and probability calibration compared with standard logistic regression, and under which data conditions differences emerge. We therefore compare PROC ENTROPY (GMED; a dual GME formulation for discrete choice) and PROC LOGISTIC across four recurring scenarios: large- n well-specified data; small- n data with strong collinearity; rare events with quasi-separation tendency; and heavy-tailed contamination/outliers. We emphasize that agreement in AUC does not necessarily imply agreement in calibrated probabilities, and we illustrate how entropy support specification can affect probability quality.

Accordingly, this study compares PROC ENTROPY (GMED) and PROC LOGISTIC across four representative data scenarios that reflect common analytical challenges encountered in practice: (1) large well-specified samples, (2) small samples with severe multicollinearity, (3) rare events with quasi-separation, and (4) heavy-tailed contamination. Emphasis is placed on whether agreement in discrimination translates into agreement in probability calibration and how entropy support specifications influence predictive performance.

2. Materials and Methods

2.1. Simulation Design

We conducted a Monte Carlo simulation comparing generalized maximum entropy estimation in SAS PROC ENTROPY (GMED) with maximum-likelihood estimation in SAS PROC LOGISTIC for a binary outcome with eight predictors (x_1 – x_8). Regression coefficients were fixed across scenarios at (0.50, −0.40, 0.30, 0.30, 0.20, −0.20, 0.10, 0.10), while the intercept varied by scenario to control event prevalence. For each simulation condition, $R = 200$ Monte Carlo replications were generated using a fixed base pseudo-random number seed of 424,242, with replication-specific seeds derived from the base seed to ensure reproducibility and independence across simulated datasets. Within each replication, observations were randomly partitioned into training (70%) and test (30%) sets. PROC LOGISTIC and PROC ENTROPY were fitted to the same training dataset and evaluated on the corresponding test dataset, permitting direct paired comparisons between the two estimation methods.

2.2. Scenario-Specific Data Generation

Predictors were generated as independent standard normal random variables unless otherwise specified. Binary outcomes were generated according to the logistic model described in Section 2.1 using the scenario-specific intercept and the fixed regression coefficients.

In Scenario 2 (S2), severe multicollinearity was induced by generating two highly correlated predictors with a target correlation of $\rho = 0.95$, using $x_1 = z$ and $x_2 = \rho z + \sqrt{1 - \rho^2}e$, where z and e are independent standard normal random variables.

In Scenario 3 (S3), rare events were generated by setting the intercept to $\beta_0 = -3.50$, resulting in a low event prevalence. A quasi-separation tendency was introduced by adding 12 to the first predictor (x_1) for a randomly selected 1% of observations, creating a small subgroup with an extremely high probability of the outcome.

In Scenario 4 (S4), heavy-tailed contamination was introduced by randomly selecting 5% of observations and perturbing a randomly selected half of the predictor variables with an additive $2 \times t(1)$ random variate, where

$t(1)$ denotes a Student's t distribution with one degree of freedom. This design generated extreme predictor values while preserving the underlying regression structure for most observations (Table 1).

Table 1. Baseline simulation scenarios and design features.

Scenario	Sample Size (n)	Design Feature	Intercept Setting
S1	5000	Large-n, well-specified predictors	$\beta_0 = -0.20$
S2	100	Small-n with high collinearity ($\text{corr}(x_1, x_2) \approx 0.95$)	$\beta_0 = -0.20$
S3	1000	Rare events with quasi-separation tendency (1%: x_1 shifted by +12)	$\beta_0 = -3.50$
S4	1000	Heavy-tailed contamination/outliers (5% perturbed with $t(1)$ noise)	$\beta_0 = 0.00$

Note. Slopes were held constant across scenarios for eight predictors (x_1 – x_8): (0.50, −0.40, 0.30, 0.30, 0.20, −0.20, 0.10, 0.10).

2.3. Estimation Procedures

Logistic regression models were fitted using PROC LOGISTIC with maximum likelihood, and predicted probabilities were obtained for the test set using the SCORE statement. GME models were fitted using PROC ENTROPY with the GMED option (TECH = NRA) and an explicit constant regressor (const = 1) with NOINT in the MODEL statement.

For GMED, error supports were specified as a 7-point symmetric grid:

$$(-0.10, -0.0667, -0.0333, 0, 0.0333, 0.0667, 0.10)$$

The ESUPPORTS option specifies the finite support points used to represent the disturbance distribution in the generalized maximum entropy model. A symmetric seven-point grid centered at zero was selected a priori as a moderate baseline specification to provide a balanced representation of uncertainty while avoiding unnecessarily restrictive or overly diffuse support ranges. Although no universally optimal support specification exists, symmetric support grids are commonly used in entropy-based estimation because they provide a flexible representation of random error without imposing directional bias (Golan et al., 1996). Coefficient supports and prior probabilities were left at the SAS default settings.

To evaluate the robustness of the GMED estimates to the choice of disturbance supports, sensitivity analyses were conducted using narrower and wider ESUPPORTS grids while keeping all other simulation settings unchanged. These analyses were designed to assess the influence of support specification on predictive performance, particularly probability calibration.

2.4. Performance Metrics

Out-of-sample predictive performance was evaluated using measures of discrimination, overall probabilistic accuracy, and calibration. Discrimination was assessed using the area under the receiver operating characteristic (ROC) curve (AUC). Overall probabilistic accuracy was evaluated using the Brier score, calculated as the mean squared error between predicted probabilities and observed outcomes, and log loss, calculated as the mean negative log-likelihood of the predicted probabilities.

To avoid numerical instability when computing log loss and calibration statistics, predicted probabilities were truncated to the interval $[10^{-12}, 1 - 10^{-12}]$.

Calibration was assessed on the test dataset using logistic recalibration, in which the observed outcome was regressed on the logit of the predicted probability:

$$\text{logit}\{P(Y = 1)\} = \alpha + \beta \text{logit}\{\hat{P}(Y = 1)\}$$

where α represents the calibration intercept (calibration-in-the-large), reflecting systematic over- or underestimation of risk, and β represents the calibration slope, which evaluates whether predicted probabilities are appropriately scaled. An ideal model has a calibration intercept of 0 and a calibration slope of 1.

2.5. Summary Measures and Paired Contrasts

For each simulation scenario and estimation method, performance metrics were summarized across $R = 200$ Monte Carlo replications as the mean and standard deviation (SD). Because PROC ENTROPY and PROC LOGISTIC were evaluated on the same simulated training and test datasets within each replication, paired comparisons were performed by calculating the within-replication difference for each performance metric:

$$\Delta = \text{ENTROPY} - \text{LOGISTIC}.$$

Positive and negative values of Δ were interpreted relative to the performance metric being evaluated. For AUC and calibration slope, positive Δ values indicate larger values for PROC ENTROPY, whereas for Brier score and log loss, negative Δ values indicate better predictive performance because lower values are preferred. Paired differences were summarized as the mean (SD), and two-sided 95% confidence intervals were computed using the Student's *t* distribution with $R - 1 = 199$ degrees of freedom.

2.6. Sensitivity Analyses

To evaluate the robustness of the primary findings, three pre-specified sensitivity analyses were conducted (Supplementary Tables S1–S6). First, the effect of event prevalence was examined in the rare-event scenario by varying the intercept (β_0) to produce approximate event rates of 1% ($\beta_0 = -4.6$), 3–5% ($\beta_0 = -3.5$), and 10–13% ($\beta_0 = -2.2$). Second, the influence of the GMED disturbance support specification (ESUPPORTS) was evaluated by comparing narrow, baseline, and wide seven-point symmetric support grids while holding all other simulation settings constant. Third, the effect of sample size under severe multicollinearity was investigated by varying the sample size ($n = 50, 100$, and 250) while maintaining a target predictor correlation of $\rho = 0.95$. These analyses were designed to evaluate the stability of discrimination, probabilistic accuracy, and calibration under varying data conditions and model specifications.

3. Results

3.1. Baseline Scenarios

Table 2 summarizes mean (SD) performance across $R = 200$ replications in the four baseline scenarios, and Figures 1–4 visualize method comparisons. In S1 (large- n), AUC and scoring rules were essentially identical, while PROC ENTROPY produced a modestly larger calibration slope ($\Delta \approx +0.024$). In S2 (small- n with collinearity), both methods exhibited degraded calibration (slopes ≈ 0.37 – 0.38), with negligible AUC differences; PROC ENTROPY showed slightly better Brier/log loss on average. In S3 (rare events with quasi-separation tendency), AUC remained similar ($\Delta\text{AUC} \approx +0.0014$), but calibration diverged substantially: PROC ENTROPY had higher calibration intercept ($\Delta \approx +0.456$) and slope ($\Delta \approx +0.278$). In S4 (heavy tails/outliers), PROC LOGISTIC showed better discrimination ($\Delta\text{AUC} \approx -0.0055$) and slightly better scoring rules, while PROC ENTROPY again produced larger calibration slopes ($\Delta \approx +0.146$).



Figure 1. Comparison of discrimination (AUC) between PROC ENTROPY (GMED) and PROC LOGISTIC across the four baseline simulation scenarios. Points represent the mean AUC across 200 Monte Carlo replications, and error bars denote 95% confidence intervals. Discrimination was similar across most scenarios, with only modest differences observed under heavy-tailed contamination (S4).

Table 2. Out-of-sample performance by scenario and method (mean (SD), R = 200).

Scenario	Method	AUC	Brier	LogLoss	Cal Intercept	Cal Slope
S1	ENTROPY	0.707 (0.014)	0.2160 (0.0042)	0.6209 (0.0094)	0.003 (0.067)	1.021 (0.101)
S1	LOGISTIC	0.707 (0.014)	0.2160 (0.0043)	0.6209 (0.0096)	0.002 (0.067)	0.997 (0.099)
S2	ENTROPY	0.591 (0.111)	0.2640 (0.0430)	0.7449 (0.1136)	−0.151 (0.463)	0.383 (0.548)
S2	LOGISTIC	0.591 (0.111)	0.2654 (0.0440)	0.7507 (0.1186)	−0.151 (0.462)	0.372 (0.534)
S3	ENTROPY	0.751 (0.071)	0.0373 (0.0100)	0.1589 (0.0330)	−0.040 (0.934)	1.128 (0.360)
S3	LOGISTIC	0.750 (0.072)	0.0374 (0.0104)	0.1600 (0.0391)	−0.496 (0.795)	0.850 (0.286)
S4	ENTROPY	0.707 (0.033)	0.2174 (0.0095)	0.6226 (0.0216)	0.001 (0.158)	1.057 (0.266)
S4	LOGISTIC	0.712 (0.030)	0.2160 (0.0100)	0.6196 (0.0236)	0.003 (0.152)	0.911 (0.198)

Note. Calibration intercept and slope are from logistic recalibration on test-set predictions: $\text{logit}\{P(Y = 1)\} = \alpha + \beta \text{logit}\{\hat{P}(Y = 1)\}$.

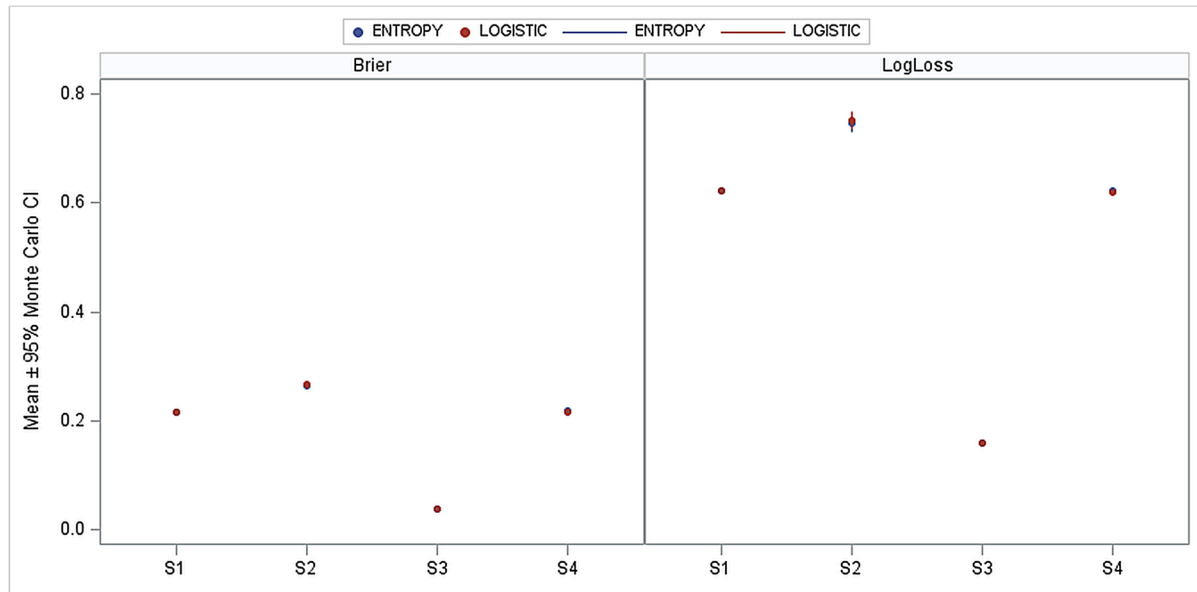


Figure 2. Comparison of probabilistic accuracy between PROC ENTROPY (GMED) and PROC LOGISTIC across the four baseline simulation scenarios. The **left panel** shows mean Brier scores and the **right panel** shows mean log loss values across 200 Monte Carlo replications. Error bars represent 95% confidence intervals. Lower values indicate better predictive performance.

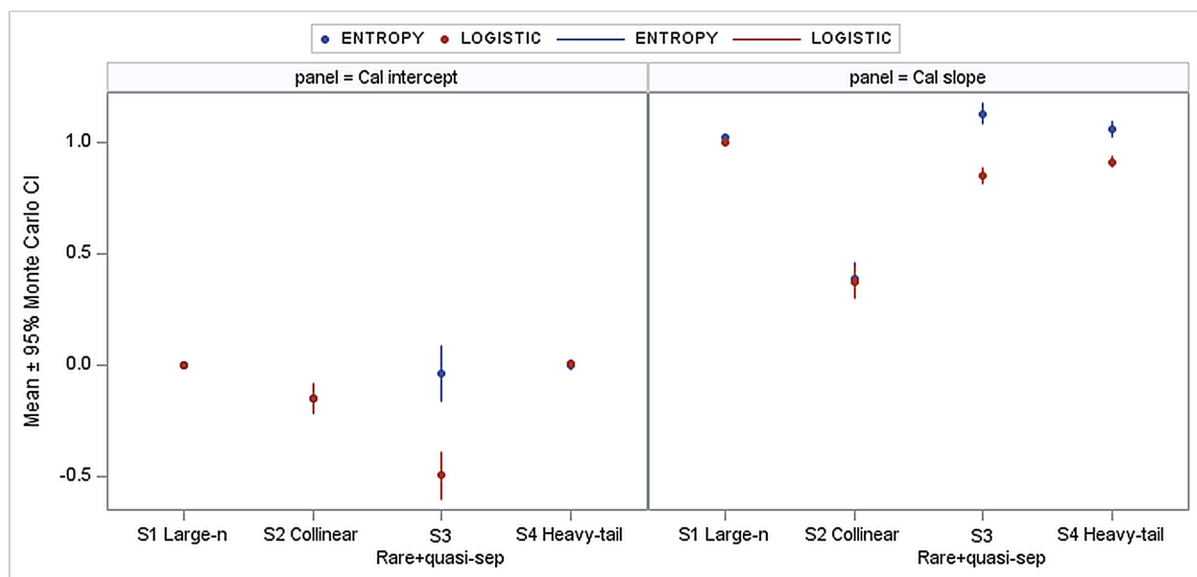


Figure 3. Comparison of calibration performance between PROC ENTROPY (GMED) and PROC LOGISTIC across the four baseline simulation scenarios. The **left panel** shows calibration intercepts (ideal = 0), and the **right panel** shows calibration slopes (ideal = 1). Points represent mean values across 200 Monte Carlo replications with 95% confidence intervals. The largest calibration differences occurred in the rare-event scenario (S3).

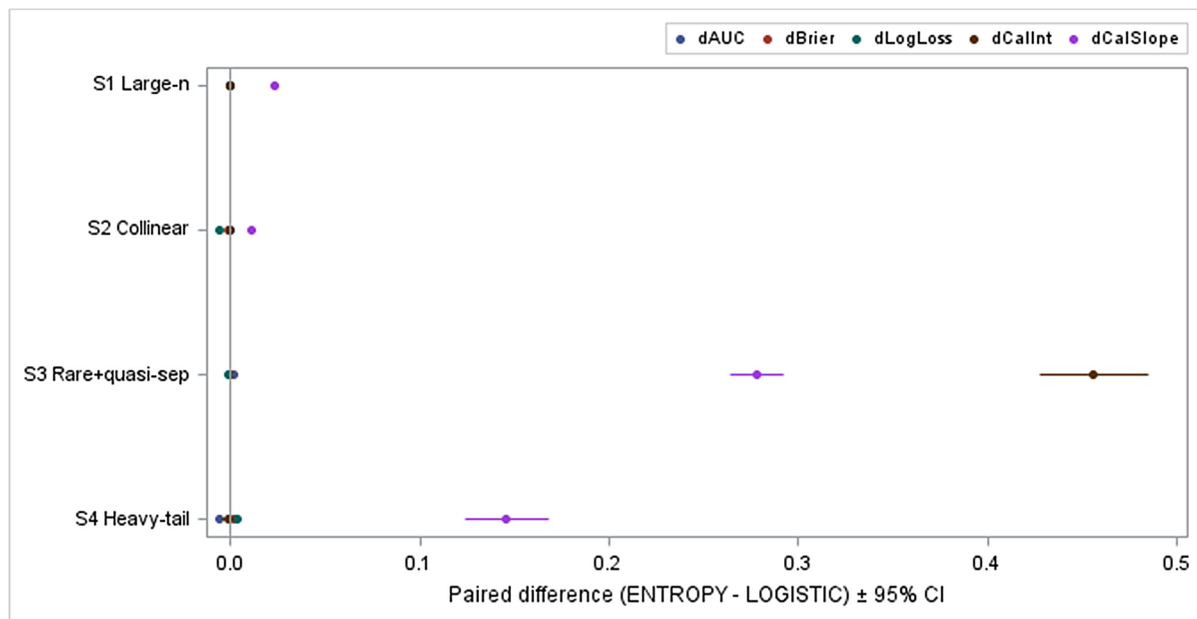


Figure 4. Paired differences in predictive performance (PROC ENTROPY – PROC LOGISTIC) across the four baseline simulation scenarios. Points represent the mean paired difference across 200 Monte Carlo replications, and error bars indicate 95% confidence intervals. Positive values indicate larger values for PROC ENTROPY, whereas negative values indicate lower values for PROC ENTROPY. Interpretation depends on the performance metric (e.g., lower Brier score and log loss indicate better performance).

3.2. Paired Differences

Table 3 reports paired differences (ENTROPY – LOGISTIC) with 95% confidence intervals. Across S1–S2, Δ AUC and scoring-rule differences were near zero, while Δ calibration slope was consistently positive. In S3, calibration differences were large despite small Δ AUC, underscoring that discrimination agreement does not guarantee calibration agreement. In S4, PROC LOGISTIC outperformed PROC ENTROPY on discrimination and scoring rules.

Table 3. Paired differences (Δ = ENTROPY – LOGISTIC), mean (SD) and 95% CI (R = 200).

Scenario	Δ AUC [95% CI]	Δ Brier [95% CI]	Δ LogLoss [95% CI]	Δ Cal Intercept [95% CI]	Δ Cal Slope [95% CI]
S1	−0.0000004 [−0.0000021, 0.0000012]	−0.0000015 [−0.0000147, 0.0000117]	−0.0000075 [−0.0000441, 0.0000291]	0.00005 [0.00004, 0.00006]	0.02369 [0.02341, 0.02397]
S2	0.0001607 [−0.0000981, 0.0004196]	−0.0014036 [−0.0015715, −0.0012356]	−0.0057876 [−0.0066585, −0.0049166]	0.00016 [−0.00003, 0.00035]	0.01091 [0.00885, 0.01296]
S3	0.0013876 [0.0005563, 0.0022188]	−0.0000691 [−0.0001564, 0.0000182]	−0.0011677 [−0.0022397, −0.0000956]	0.45605 [0.42727, 0.48484]	0.27826 [0.26416, 0.29236]
S4	−0.0054660 [−0.0070077, −0.0039242]	0.0014043 [0.0008950, 0.0019135]	0.0030283 [0.0016692, 0.0043874]	−0.00167 [−0.00489, 0.00155]	0.14569 [0.12356, 0.16781]

Note. Negative Δ indicates lower (better) values for Brier and log loss for ENTROPY relative to LOGISTIC; positive Δ indicates higher values. For AUC and calibration slope, positive Δ indicates larger values for ENTROPY.

3.3. Sensitivity Analyses

Sensitivity analyses (Supplement Tables S1–S6) showed that calibration differences were most sensitive to event prevalence and to the ESUPPORTS width used in PROC ENTROPY. As prevalence decreased, differences in calibration intercept and slope increased. Wider ESUPPORTS values tended to increase calibration divergence and could worsen scoring rules, whereas narrow ESUPPORTS reduced divergence. Varying sample size in the collinearity setting showed that both procedures stabilized as an increased, with shrinking differences in scoring rules and calibration.

4. Discussion

This Monte Carlo study compared PROC LOGISTIC (maximum likelihood) with PROC ENTROPY (GMED) across four baseline data scenarios and three sensitivity families. In large samples with well-specified predictors (S1), the two procedures were nearly indistinguishable in discrimination and scoring rules, consistent with the expectation that entropy-based estimates can align closely with likelihood-based estimates when information is plentiful. Nevertheless, calibration slopes differed systematically, indicating small but consistent differences in the scaling of predicted probabilities.

An important observation from the present study is that similar discrimination does not necessarily imply similar probability estimation. Discrimination metrics such as the AUC evaluate a model's ability to correctly rank individuals according to risk, whereas calibration evaluates whether predicted probabilities correspond to observed event frequencies. Consequently, two models may achieve nearly identical AUC values while producing systematically different probability estimates. The larger calibration slopes observed for PROC ENTROPY under rare-event conditions may reflect the way generalized maximum entropy distributes limited information across the coefficient support space subject to entropy constraints. Unlike maximum-likelihood estimation, coefficient scaling depends partly on the specified support structure, which can influence probability scaling even when patient ranking remains largely unchanged.

Conceptually, generalized maximum entropy and maximum-likelihood estimation approach parameter estimation from different perspectives. Maximum-likelihood estimation identifies parameter values that maximize the observed likelihood, whereas generalized maximum entropy estimates coefficients by maximizing entropy subject to moment constraints imposed by the observed data. Because coefficient estimates are represented through finite support points with associated probability weights, the resulting coefficient scaling may depend partly on the specified support structure and the available information. Under well-behaved data conditions, both estimation approaches tend to produce similar predictive performance. However, when information becomes limited—as in rare-event settings or near-separation—the entropy constraints may influence the scaling of predicted probabilities while preserving much of the relative ordering of observations. This provides one plausible explanation for the observed pattern of similar discrimination but different probability calibration in the present simulations.

In the small-sample collinearity setting (S2), both procedures exhibited substantial calibration degradation, emphasizing that unstable probability estimation can persist even when AUC differences are negligible. The rare-event scenario (S3) produced the clearest divergence, with similar discrimination but substantially different calibration, demonstrating that discrimination alone is insufficient for evaluating probability estimation. The distinct findings under heavy-tailed contamination are discussed separately below.

A different pattern emerged under heavy-tailed contamination. In this setting, PROC LOGISTIC demonstrated slightly better discrimination and scoring-rule performance, whereas PROC ENTROPY continued to produce systematically larger calibration slopes. Heavy-tailed contamination introduces extreme predictor values that may influence the allocation of probability mass within the entropy framework, thereby affecting probability scaling even when overall discrimination remains similar. These findings emphasize that calibration should be routinely evaluated whenever entropy-based estimation is used for predictive modeling, particularly in datasets containing rare events or influential observations.

Sensitivity analyses reinforced two practical observations. First, calibration differences increased as event prevalence decreased. Second, probability estimation with PROC ENTROPY was sensitive to the choice of ESUPPORTS, indicating that the support specification functions as a tuning parameter governing the flexibility of the entropy model. Narrower support grids generally produced calibration performance that more closely resembled logistic regression, whereas wider support grids resulted in larger calibration divergence despite similar discrimination. Therefore, analysts should routinely evaluate alternative ESUPPORTS specifications, report calibration diagnostics, and consider post hoc calibration when accurate probability estimation is important.

From a methodological perspective, calibration differences become important whenever predicted probabilities are used quantitatively rather than solely for ranking individuals. Two models with nearly identical AUC values may produce different absolute risk estimates because discrimination assesses relative ordering, whereas calibration evaluates the agreement between predicted and observed event probabilities. Consequently, differences in calibration may alter risk classification, influence clinical decision thresholds, and affect downstream decisions such as treatment selection, patient monitoring, or resource allocation. These findings reinforce current recommendations that discrimination metrics should be complemented by calibration assessment whenever predictive models are intended to support clinical or policy decisions.

Although this study focused on comparing generalized maximum entropy estimation with conventional maximum-likelihood logistic regression, several alternative approaches including ridge and lasso penalization,

Firth bias-reduced logistic regression, and Bayesian logistic regression have also been developed to address small samples, multicollinearity, rare events, and separation. These methods differ in their underlying estimation principles and assumptions and therefore address somewhat different methodological objectives. The present study was designed to evaluate the practical agreement between likelihood-based and entropy-based estimation rather than to identify an optimal estimator among all available approaches. Future studies comparing these methods using common predictive performance metrics, including probability calibration, would further inform methodological guidance for applied researchers.

This study has several strengths, including a comprehensive Monte Carlo design, multiple realistic data scenarios, evaluation of both discrimination and calibration, and extensive sensitivity analyses. Nevertheless, some limitations should be acknowledged. The simulations focused on the SAS implementations of PROC LOGISTIC and PROC ENTROPY and did not compare generalized maximum entropy with other alternative approaches such as ridge or lasso penalization, Firth bias-reduced logistic regression, or Bayesian logistic regression. Future research should evaluate these methods using similar calibration-focused performance metrics and real-world datasets.

Although post hoc recalibration methods (e.g., recalibration intercept and slope adjustment) may reduce some of these differences, the objective of the present study was to compare the native predictive behavior of PROC LOGISTIC and PROC ENTROPY rather than optimized prediction pipelines. Evaluation of recalibration strategies represents an important direction for future methodological research.

Based on the present findings, we suggest several practical guidelines for analysts using PROC ENTROPY. First, a symmetric baseline support grid provides a reasonable starting point for most applications. Second, support specifications should not be viewed as fixed defaults but rather as tuning parameters that may influence probability calibration. Third, when prediction or risk estimation is the primary objective, sensitivity analyses using narrower and wider ESUPPORTS values should be routinely performed to assess the robustness of calibration. Finally, calibration diagnostics should be reported alongside discrimination measures whenever entropy-based models are used for probability estimation.

Overall, the findings demonstrate that agreement in discrimination should not be interpreted as agreement in probability estimation. Rather, agreement should be viewed as practical similarity across multiple aspects of predictive performance, including discrimination, probabilistic accuracy, and calibration. For researchers using generalized maximum entropy methods, routine evaluation of calibration and sensitivity to support specifications should be considered essential components of predictive model assessment.

5. Conclusions

The principal finding of this study is that PROC ENTROPY (GMED) and PROC LOGISTIC frequently demonstrated similar discrimination and overall predictive accuracy, yet probability calibration was more sensitive to the underlying data conditions and entropy support specification.

From a practical perspective, researchers should recognize that models with comparable discrimination may nevertheless produce substantially different probability estimates. Consequently, routine assessment of calibration should accompany discrimination metrics whenever generalized maximum entropy methods are applied to predictive modeling.

The simulation results further suggest that the information-theoretic constraints underlying generalized maximum entropy estimation may influence probability scaling even when discrimination remains largely unchanged, highlighting the importance of evaluating calibration whenever entropy-based methods are used for predictive modeling.

Accordingly, “agreement” between the two methods should be interpreted as practical similarity across multiple aspects of predictive performance rather than exact numerical equivalence. Agreement in discrimination should not be interpreted as evidence of agreement in probability calibration.

Supplementary Materials

The additional data and information can be downloaded at: <https://jmasm.com/index.php/jmasm/article/view/1496/s1>.

Funding

This research received no external funding.

Institutional Review Board Statement

Ethical review and approval were waived because this study involved computer simulation methods only and did not involve human participants, animals, or identifiable human data.

Informed Consent Statement

Not applicable.

Data Availability Statement

No human participant data were generated or analyzed in this study. The simulation code, parameter settings, and generated datasets supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

Use of AI and AI-Assisted Technologies

ChatGPT (OpenAI) was used for English-language editing and formatting assistance. The author takes full responsibility for the final text, analyses, and conclusions.

References

1. Agresti, A. *Categorical Data Analysis*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2013.
2. Hosmer, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2013.
3. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather. Rev.* **1950**, *78*, 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
4. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. <https://doi.org/10.1148/radiology.143.1.7063747>.
5. Steyerberg, E.W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 2nd ed.; Springer Cham, Switzerland, 2019.
6. Van Calster, B.; McLernon, D.J.; van Smeden, M.; et al. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. <https://doi.org/10.1186/s12916-019-1466-7>
7. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **1993**, *80*, 27–38. <https://doi.org/10.1093/biomet/80.1.27>.
8. King, G.; Zeng, L. Logistic regression in rare events data. *Political Anal.* **2001**, *9*, 137–163. <https://doi.org/10.1093/oxfordjournals.pan.a004868>.
9. Albert, A.; Anderson, J.A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **1984**, *71*, 1–10. <https://doi.org/10.1093/biomet/71.1.1>.
10. Heinze, G.; Schemper, M. A solution to the problem of separation in logistic regression. *Stat. Med.* **2002**, *21*, 2409–2419. <https://doi.org/10.1002/sim.1047>.
11. Le Cessie, S.; van Houwelingen, J.C. Ridge estimators in logistic regression. *J. R. Stat. Soc. Ser. C* **1992**, *41*, 191–201. <https://doi.org/10.2307/2347628>.
12. Pregibon, D. Logistic regression diagnostics. *Ann. Stat.* **1981**, *9*, 705–724. <https://doi.org/10.1214/aos/1176345513>.
13. Golan, A.; Judge, G.G.; Miller, D. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*; John Wiley & Sons: New York, NY, USA, 1996.
14. Golan, A. *Foundations of Information-Theoretic Methods in Econometrics*; Oxford University Press: Oxford, UK, 2018.