Received: 6 December 2018 | Accepted: 4 September 2019

DOI: 10.1111/biom.13249

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

BIOMETRIC PRACTICE

Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test

Giovanni Nattino[1,2] | **Michael L. Pennell**[2] | **Stanley Lemeshow**[2]

[1]The Ohio Colleges of Medicine Government Resource Center, Ohio State University, Columbus, Ohio

[2]Division of Biostatistics, College of Public Health, Ohio State University, Columbus, Ohio

Correspondence
Giovanni Nattino, The Ohio Colleges of Medicine Government Resource Center, Ohio State University, Columbus, OH 43210.
Email: nattino.1@osu.edu

Abstract

Evaluating the goodness of fit of logistic regression models is crucial to ensure the accuracy of the estimated probabilities. Unfortunately, such evaluation is problematic in large samples. Because the power of traditional goodness of fit tests increases with the sample size, practically irrelevant discrepancies between estimated and true probabilities are increasingly likely to cause the rejection of the hypothesis of perfect fit in larger and larger samples. This phenomenon has been widely documented for popular goodness of fit tests, such as the Hosmer-Lemeshow test. To address this limitation, we propose a modification of the Hosmer-Lemeshow approach. By standardizing the noncentrality parameter that characterizes the alternative distribution of the Hosmer-Lemeshow statistic, we introduce a parameter that measures the goodness of fit of a model but does not depend on the sample size. We provide the methodology to estimate this parameter and construct confidence intervals for it. Finally, we propose a formal statistical test to rigorously assess whether the fit of a model, albeit not perfect, is acceptable for practical purposes. The proposed method is compared in a simulation study with a competing modification of the Hosmer-Lemeshow test, based on repeated subsampling. We provide a step-by-step illustration of our method using a model for postneonatal mortality developed in a large cohort of more than 300 000 observations.

KEYWORDS

calibration, goodness of fit, Hosmer-Lemeshow test, large samples, logistic regression, noncentrality parameter

1 | INTRODUCTION

Statistical models estimating the probability of binary outcomes are routinely used in a variety of research fields. Among the available methods to generate such models, logistic regression is the most popular. The performance of a model is usually evaluated under two dimensions: discrimination and calibration (Hosmer *et al.*, 2013). The discrimination of a model is its capability of assigning higher probabilities of the outcome to those observations that actually experience the outcome. A well-established measure of discrimination is the area under the receiver operating characteristic (ROC) curve. The calibration of a model quantifies the accuracy of the estimated probabilities of the outcome. Several tests and graphical methods have been proposed to assess the calibration of a model, which is often referred to as "goodness of fit." Among the goodness of fit tests, the Hosmer-Lemeshow (HL) test is the most widely applied approach. The idea of the test is to partition the observations into groups and construct a chi-squared statistic that summarizes the discrepancy between the number of observed and expected events within all combinations of group and outcome state.

As with most goodness of fit tests, the HL test is designed to decide between a null hypothesis of perfect fit, where the probabilities assumed by the model are hypothesized to coincide with the real probabilities, and a general alternative hypothesis of nonperfect fit. Albeit reasonable, such a formulation of the statistical hypotheses is problematic in large samples. Because the power of goodness of fit tests increases with the sample size, minuscule discrepancies between a model's estimates and actual probabilities are likely to cause the rejection of the null hypothesis with large sample sizes, even if such discrepancies are irrelevant for the scope of the model. Over the past two decades, several studies have documented the frequent rejection of acceptable—though nonperfect—models (eg, Kramer and Zimmerman, 2007; Paul *et al.*, 2013). This phenomenon has limited the usefulness of the HL test in practice, where researchers rarely deal with the true model and a certain degree of model misspecification is inevitable.

Different approaches have been proposed to overcome the limitations of goodness of fit tests in large samples. Some studies have suggested the assessment of a model's calibration with graphical methods (Austin and Steyerberg, 2014; Moons *et al.*, 2015; Nattino *et al.*, 2016). The idea is to represent the relationship between a model's predictions (on the x-axis) and outcome rates (on the y-axis) and evaluate the goodness of fit by looking at the closeness of the resulting relationship to the 45° line. Although this type of assessment does not suffer from the shortcomings of goodness of fit tests, it is highly subjective and does not allow rigorous comparisons of competing models.

Other studies have proposed modifications of the HL testing procedure to standardize its power across sample sizes, with the goal of limiting the excessively high rejection rates of the test in large samples. Paul *et al.* (2013) studied the relationship between the number of groups formed for the HL statistic and the power of the test, providing recommendations on how to choose the number of groups to attain constant power at different sample sizes. Given a model to be evaluated and a sample of size n, Yu *et al.* (2017) proposed an alternative procedure to estimate the distribution of the HL statistic in a "standard" sample of size $m < n$. The goodness of fit of the model is assessed by comparing a random draw from this standardized distribution with an appropriate critical value. Lai and Liu (2018) recently described a similar procedure, where the distribution of the HL statistic is estimated using numerous subsamples of size m. As in the procedure of Yu *et al.* (2017), the decision about the fit of the model is made by comparing a random draw from the estimated distribution with an appropriate critical value. Interestingly, Lai and Liu (2018) compared these three methods in a simulation study and found that only their approach had constant rejection rates across sample sizes.

Although the procedures proposed by Yu *et al.* (2017) and Lai and Liu (2018) were designed to decide whether the fit of a model is acceptable, they are not structured as traditional hypothesis tests. The responses of the procedures are binary (the goodness of fit is rejected or not) and the decision is left to chance, because it is based on a random draw. In particular, different researchers, applying the same model to the same data, might reach different conclusions about the quality of the model. This is a major limitation of these approaches.

In this paper, we address the shortcomings of the HL test in large samples from a different perspective. First, we identify a measure of goodness of fit that is independent of sample size. This measure is based on the noncentrality parameter that characterizes the noncentral chi-squared distribution of the HL test. This idea has been explored in the assessment of structural equation models but is novel in the framework of logistic regression (Steiger, 1990; Browne and Cudeck, 1992; Steiger and Fouladi, 1997). Then, taking advantage of this measure of goodness of fit, we propose a statistical test that considers a more conservative null hypothesis, including models with an acceptable, albeit not perfect, fit. Because the null hypothesis of the proposed test comprises models with a satisfactory fit, our approach can be used to discern between acceptable and poorly fitting models in large samples. Importantly, the HL test is a special case of the proposed test.

The remainder of this paper is organized as follows. Section 2 provides the background on the HL test. We introduce the proposed measure of goodness of fit and the related statistical test in Section 3. Section 4 describes the simulation study that evaluates the proposed method and compares it to the resampling procedure described by Lai and Liu (2018). We apply our method to a model predicting postneonatal mortality in a sample of more than 300 000 observations in Section 5. We discuss the limitations and strengths of the proposed method in Section 6.

2 | BACKGROUND: THE HOSMER-LEMESHOW GOODNESS OF FIT TEST

Given a set of predictors, $X_1, X_2, \dots, X_p$, and a binary outcome, Y, a logistic regression model has form $\text{logit}\{P(Y=1|X_1, X_2, \dots, X_p)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$, where the logit link is defined as $\text{logit}(p) = \ln(p/(1-p))$. The parameters are estimated by numerically maximizing the likelihood of the model given a sample $\{(Y_i, X_{i1}, X_{i2}, \dots, X_{ip})\}_{i=1,\dots,n}$. Denote the probability assumed by the model for subject i with $p_i = \{1 + \exp(-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}))\}^{-1}$.

Plugging the estimates $\hat{\beta}_0,\ \hat{\beta}_1,\ \dots,\hat{\beta}_p$ into this equation, it is possible to estimate the probability of the positive outcome (ie, $Y = 1$) for each subject in the sample.

A model is perfectly calibrated if the true probabilities of the outcome, $\{\pi_i = P(Y_i = 1|X_{i1}, X_{i2}, \dots,\ X_{ip})\}_{i\ =\ 1,\dots,n}$, coincide with the model's estimates. We refer to $H_0\ :\ \pi_i = p_i$ for all $i = 1,\dots,n$ as the hypothesis of perfect fit, which is the null hypothesis of the HL test. The alternative hypothesis is $H_a\ :\ \pi_i \neq p_i$ for some $i = 1,\dots,n$. Note that the sample size, n, is considered as a fixed parameter in goodness of fit settings, as it is the dimension of the parameters involved in the hypotheses. Importantly, H_0 does not necessarily imply that a model coincides with the data-generating mechanism, that is, $Y_i \sim \text{Bernoulli}(p_i)$. A model may be well calibrated even if it omits important predictors as long as it describes the conditional probabilities π_i correctly.

To compute the HL statistic, observations are ordered and divided into G groups. For each group g, denote the frequency of events and nonevents with O_{1g} and O_{0g} and the number of events and nonevents predicted by the model with E_{1g} and E_{0g}. Constructing 10 approximately equinumerous groups is a common choice in practice. The statistic summarizes the discrepancies between observed and expected number of events and nonevents over all G groups:

$$\hat{C} = \sum_{g=1}^{G} \left(\frac{\left(O_{1g} - E_{1g}\right)^2}{E_{1g}} + \frac{\left(O_{0g} - E_{0g}\right)^2}{E_{0g}} \right). \tag{1}$$

Under the null hypothesis of perfect fit, the statistic is asymptotically distributed as a chi-squared random variable with $G-2$ degrees of freedom (Hosmer and Lemesbow, 1980). In this paper, we focus on the in-sample assessment of the goodness of fit; that is, the case where the model is evaluated on the sample used to fit the model. However, our results promptly extend to the setting of external validation; that is, when the model under assessment is applied to independent data. In this case, the degrees of freedom of the statistic are G(Hosmer *et al.*, 2013).

If the null hypothesis does not hold, the distribution of the statistic is noncentral chi-squared with the same degrees of freedom and noncentrality parameter λ (Moore and Spruill, 1975). The larger the discrepancy between a model's estimates and true probabilities of the outcome, the larger the λ. When the model fits the data perfectly, $\lambda = 0$ and the distribution of the statistic is the central chi-squared distribution. Importantly, λ also depends on the sample size. If a model is fit on samples of increasing size taken from the same population, λ increases linearly with the sample size (ie, $\lambda \propto n$). This result follows from Theorem 1 of Dahiya and Gurland (1973) and is formally shown in Web Appendix A. Building on this result, we propose a measure of model goodness of fit based on the noncentrality parameter λ.

3 | MODIFYING THE HOSMER-LEMESHOW TEST

3.1 | The standardized noncentrality parameter

Although stronger departures from the perfect fit correspond to larger values of λ, its value cannot be directly used to measure the goodness of fit of a model. Because λ increases linearly with the sample size, equally calibrated models correspond to different values of λ on samples with different sizes. However, the relationship between λ and n suggests a standardization of λ that is comparable across sample sizes. We propose the following transformation of λ:

$$\epsilon = \sqrt{\frac{\lambda}{n}}. \tag{2}$$

We refer to ϵ as the standardized noncentrality parameter of the HL statistic. Holding the departure from perfect fit constant, the fact that $\lambda \propto n$ implies that ϵ is asymptotically constant. Because the noncentrality parameter λ has the scale of a chi-squared random variable, we propose taking the square root of the ratio, in order to obtain a measure that has the scale of a standard normal random variable. Notably, n is a fixed parameter in the goodness of fit framework and, therefore, ϵ is a valid target to infer about the calibration of a model.

In the context of structural equation models, Steiger (1990) proposed a similar standardization of the noncentrality parameter of the discrepancy function, a goodness of fit statistic with noncentral chi-squared distribution. The author introduced the root mean squared error of approximation (RMSEA), defined as the square root of the ratio of the noncentrality parameter over n and the degrees of freedom (Steiger and Fouladi, 1997). Such a transformation of the noncentrality parameter only differs from ϵ in Equation (2) by the presence of the degrees of freedom of the statistic in the denominator. The purpose of dividing by the degrees of freedom is to facilitate the interpretation of the ratio as the discrepancy per degree of freedom (Browne and Cudeck, 1992). In our context, because the value of λ does not have an intuitive interpretation, we prefer a simpler expression that does not involve the degrees of freedom.

Equation (2) provides the parameter that can be used to characterize the goodness of fit of a model. However, λ is unknown and must be estimated. We use the same approach suggested for the RMSEA and estimate λ with $\hat{\lambda} = \max\{\hat{C} - (G-2),\ 0\}$(Saxena and Alam, 1982). Therefore, we consider the following estimator of ϵ:

$$\hat{\epsilon} = \sqrt{\frac{\max\left\{\hat{C} - (G-2),\ 0\right\}}{n}}. \tag{3}$$

A confidence interval (CI) for ϵ can be constructed using CIs for λ. Kent and Hainsworth (1995) described and compared different approaches for generating CIs for the noncentrality parameter of noncentral chi-squared distributions. On the basis of the coverage and simplicity of the calculations, the authors recommended the use of symmetric-range CIs. This family of CIs is generated by inverting a probability interval for $\sqrt{\hat{C}}$. Assuming that λ is the true noncentrality parameter of $\hat{C}$, consider $c_L(\lambda) = \max\{\sqrt{\lambda} - b(\lambda), 0\}$ and $c_U(\lambda) = \sqrt{\lambda} + b(\lambda)$, with $b(\lambda)$ chosen to give $P\,(c_L(\lambda) \le \sqrt{\hat{C}} \le c_U(\lambda)) = 1 - \alpha$. For any given value of λ, the distribution of $\hat{C}$ is known; it is noncentral chi-squared with $G - 2$ degrees of freedom and noncentrality parameter λ. Thus, for a fixed α, we can compute $c_L(\lambda)$ and $c_U(\lambda)$ numerically. A graphical representation of $c_L(\lambda)$ and $c_U(\lambda)$ is provided in figure 1 of Kent and Hainsworth (1995) and reproduced in Figure 1 for the case $G = 10$. When the value $\hat{c}$ of the statistic is observed, the set of values λ' such that $c_L(\lambda') \le \sqrt{\hat{c}} \le c_U(\lambda')$ is a valid $(1 - \alpha)100\%$ CI for λ. The smallest and largest values of this set, namely $\lambda_L(\hat{c})$ and $\lambda_U(\hat{c})$, can be obtained by drawing a horizontal line in correspondence of $\sqrt{\hat{c}}$ in Figure 1 and identifying the intersections of this line with the curves $c_L(\lambda)$ and $c_U(\lambda)$. The CI $[\epsilon_L(\hat{c}), \epsilon_U(\hat{c})]$ for ϵ is constructed plugging the extremes of the CI for λ into Equation (2).

Similarly, using the one-sided CIs for λ, it is possible to construct one-sided CIs $[\epsilon_L(\hat{c}), \infty)$ for ϵ. The lower bound $\lambda_L(\hat{c})$ of the one-sided CI has a particularly simple expression. In the one-sided case, we have that $c_U(\lambda) = \sqrt{\chi^2_{\lambda,G-2,\alpha}}$, where $\chi^2_{\lambda,G-2,\alpha}$ denotes the α-level upper quantile of a noncentral chi-squared distribution with noncentrality parameter λ and $G - 2$ degrees of freedom. The one-sided CI of λ contains the values λ' such that $\sqrt{\hat{c}} \le c_U(\lambda')$, that is, $\sqrt{\hat{c}} \le \sqrt{\chi^2_{\lambda',G-2,\alpha}}$. The expression that identifies the lower bound $\lambda_L(\hat{c})$ depends on the value of $\hat{c}$. If $\hat{c} > \chi^2_{0,G-2,\alpha}$, $\lambda_L(\hat{c})$ can be computed solving for λ' the equation $\hat{c} = \chi^2_{\lambda',G-2,\alpha}$. If $\hat{c} \le \chi^2_{0,G-2,\alpha}$, $\lambda_L(\hat{c}) = 0$.

3.2 | A goodness of fit test using the standardized noncentrality parameter

The standardized noncentrality parameter ϵ can be used to measure the goodness of fit of a model. If the model fits the data perfectly, the $\hat{C}$statistic has a central chi-squared distribution ($\lambda = 0$) and, consequently, $\epsilon = 0$. The worse the fit, the larger the value of ϵ. Because of the standardization with respect to the sample size, the value of ϵ only depends on the departure of the model from perfect fit.

Notably, the traditional HL test is equivalent to the test generated by the inversion of the one-sided $(1 - \alpha)100\%$ CI of ϵ discussed in Section 3.1. Proposition 1 describes this equivalence and establishes a correspondence between the HL test and a test evaluating the hypotheses $H_0 : \epsilon = 0$ versus $H_a : \epsilon > 0$. The proof of the result is provided in Web Appendix A.

Proposition 1. *Consider the hypotheses $H_0 : \epsilon = 0$ versus $H_a : \epsilon > 0$ and the statistical test that rejects H_0 at level α if zero does not belong to $[\epsilon_L(\hat{c}), \infty)$, the one-sided $(1 - \alpha)$ 100% CI of ϵ discussed in Section 3.1. This test is equivalent to the HL test.*

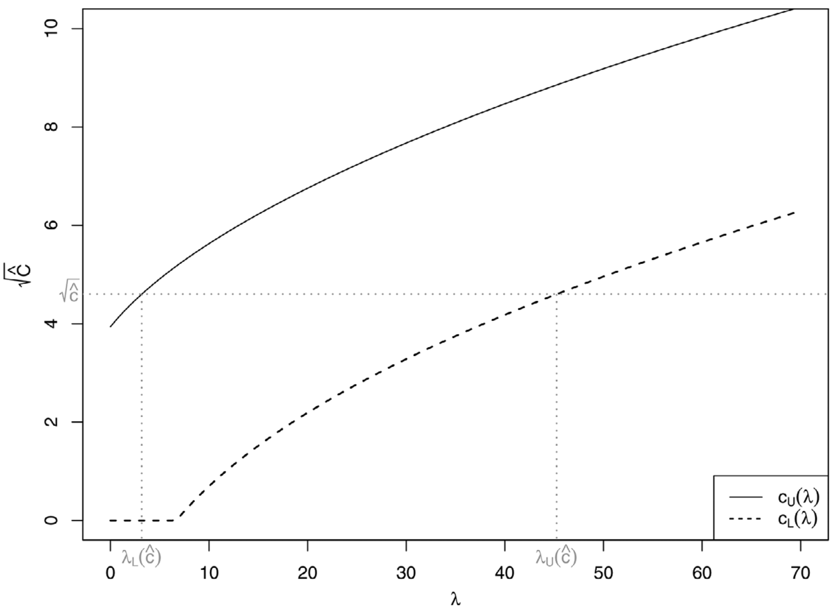
Formulating the hypotheses of the HL test on a one-dimensional parameter provides the framework to go beyond the unrealistic null hypothesis of perfect fit of goodness of fit tests. Suppose that we knew that models with ϵ smaller than a constant ϵ_0 would be acceptable for the purposes of the model under assessment. A reasonable goodness of fit test would evaluate whether the model's fit is acceptable with the hypotheses $H_0 : \epsilon \le \epsilon_0$ versus $H_a : \epsilon > \epsilon_0$. Testing whether the model's fit is perfect is the most extreme case of this family of hypotheses, with $\epsilon_0 = 0$.

By fixing the value of ϵ_0 and by inverting the one-sided CI for ϵ, we construct a test to evaluate $H_0 : \epsilon \le \epsilon_0$ versus $H_a : \epsilon > \epsilon_0$. A procedure that accepts H_0 if ϵ_0 lies inside the one-sided $(1 - \alpha)100\%$ CI of ϵ and rejects H_0 otherwise is a level α test for these hypotheses. This is shown in Proposition 2, which also provides an expression for the P-value of the test.

Proposition 2. *Consider the hypotheses $H_0 : \epsilon \le \epsilon_0$ versus $H_a : \epsilon > \epsilon_0$ and the statistical test that rejects H_0 at level α if ϵ_0 does not belong to $[\epsilon_L(\hat{c}), \infty)$, the one-sided $(1 - \alpha)$ 100% CI of ϵ discussed in Section 3.1. This is a level α test and the corresponding P-value is given by $1 - F_{\epsilon_0^2 n,\, G-2}(\hat{c})$, where $F_{\in_0^2 n, G-2}(\cdot)$ is the cumulative density function (CDF) of a noncentral chi-squared distribution with noncentrality parameter $\epsilon_0^2 n$ and $G - 2$ degrees of freedom.*

The proof is provided in Web Appendix A. So far, we have assumed that we know the value ϵ_0 characterizing models with acceptable fit. Unfortunately, the standardized noncentrality parameter ϵ is not interpretable. A possible strategy to identify a meaningful value of ϵ_0 would be to carry out extensive simulations and evaluate what values of ϵ would correspond to acceptable fit for the scope of the model under assessment. We propose an alternative path, taking advantage of the correspondence between the value of the $\hat{C}$ statistic and the estimator of ϵ (Equation (3)). We set our significance level to $\alpha = 0.05$ and we consider a model that would be flagged as borderline-significant—that is, attaining a P-value of .05—in a very large sample. We consider a sample size $n_0 = 10^6$. Knowing the high power of the traditional HL test in large samples, such a model would likely not be the correct model, though its fit would definitely be acceptable in most practical

FIGURE 1 Graphical representation of the curves $c_L(\lambda)$ and $c_U(\lambda)$ that are used to compute the confidence intervals of λ. The figure considers the case $G =$ 10. When the value $\hat{c}$ of the statistic is observed, $\lambda_L(\hat{c})$ and $\lambda_U(\hat{c})$ can be obtained by drawing an horizontal line in correspondence of $\sqrt{\hat{c}}$ and identifying the intersections of this line with the curves $c_U(\lambda)$ and $c_L(\lambda)$, respectively



applications. We can define the value of ϵ_0 for this "acceptable" model as

$$\epsilon_0 = \sqrt{\frac{\chi^2_{\lambda=0,df=G-2,\alpha=0.05} - (G-2)}{n_0}}. \tag{4}$$

For the case $G = 10$ and $n_0 = 10^6$, Equation (4) results in $\epsilon_0 = 2.74 \times 10^{-3}$. With this choice of ϵ_0, the procedure described above compares the fit of a model with the fit of a model that would be considered as borderline-significant in a sample of 1 million observations.

3.3 | Relationship with the traditional Hosmer-Lemeshow test

As described in Section 3.2, the P-value of the proposed test is computed by comparing the value of $\hat{C}$, the traditional HL statistic, to the noncentral chi-squared distribution with noncentrality parameter $\epsilon_0^2 n$ and $G-2$ degrees of freedom. In contrast, the traditional test compares the value of $\hat{C}$ to a central chi-squared distribution with $G-2$ degrees of freedom. As the two tests are based on the same statistic, there is a one-to-one relationship between the P-value of the proposed test and that of the traditional HL test. Figure 2 depicts such a relationship for different sample sizes. Note that the proposed test is always more conservative than the traditional HL test (the P-values of the proposed test are always larger than the P-values of the traditional test); this is a desired result: our goal is to limit the rejection of models with acceptable, albeit nonperfect, fit. Moreover, the difference between the proposed modification and the traditional test increases in larger samples; this is another desired property.

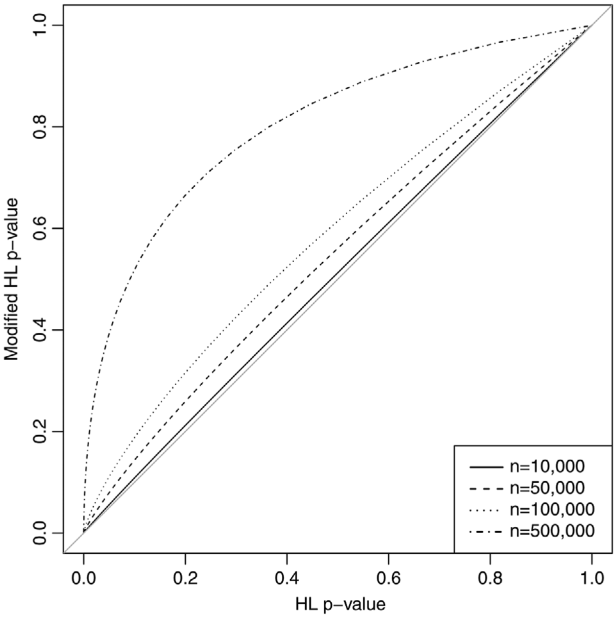


FIGURE 2 Relationship between traditional HL P-value and P-value of the proposed test for different sample sizes

For moderate sample sizes, such as n =10 000, the power of the HL test is reasonable and the rejected models are likely to be poorly calibrated. In this case, the P-value of the modified test is extremely similar to that of the traditional approach.

Thus, Figure 2 implies that the proposed testing procedure can be applied to any scenario where using the HL test would be appropriate, as the two tests rely on the same assumptions. Therefore, the proposed methodology is suited to entirely replace the HL test, regardless of the sample size.

4 | SIMULATION STUDY

4.1 | Study design

We use Monte Carlo simulations to evaluate the type-I error rate and the power of our approach. In each simulated dataset, we estimated ϵ and applied the proposed goodness of fit test, as described in Proposition 2 (with ϵ_0 fixed to the value in Equation (4), with $n_0 = 10^6$). The proposed method was compared to the traditional HL test, described in Section 2, and the resampling approach proposed by Lai and Liu (2018). In the Lai and Liu approach, one randomly generates M subsamples of size m and, in each subsample, fits the model under assessment and computes the $\hat{C}$ statistic. The null hypothesis of perfect fit is rejected with probability equal to the upper tail cut by $\chi^2_{0,G-2,0.05}$ under the empirical distribution of the computed $\hat{C}$ statistics. We implemented the Lai and Liu approach using $M = 1000$ subsamples of size $m = 1000$, as recommended by the authors. We considered $G = 10$ groups in all of the methods. For each test, we evaluated the rejection rate at the $\alpha = 0.05$ level.

Table 1 outlines the scenarios considered. We considered three families of simulations. In the first set of simulations, we generated probabilities using a second-order polynomial model with form logit $(p_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2$, where the values X_{i1} were sampled from a standard normal distribution. The outcome values were sampled as Bernoulli random variables with probabilities p_i. The data generated were used to fit a model that did not include the quadratic term. The true values of the coefficients are reported in Table 1. We fixed $\beta_1 = 1$ and we chose three different values for the intercept, to consider outcome rates of (approximately) 1%, 10%, and 20%. For each outcome rate, we identified three values of β_2: the case where the fitted model was the correct model, that is, $\beta_2 = 0$, and two other values of the parameter, chosen so that the differences between true probabilities and predictions from the misspecified model were small and large, respectively. Because ϵ can be considered as a measure of goodness of fit, we tuned β_2 to attain similar values of $\hat{\epsilon}$ across simulations with different outcome rates. In particular, for the small departure models, β_2 was selected to attain an average value of $\hat{\epsilon}$ of ϵ_0 (as defined in Equation (4)) in samples of 1 million records. For the scenarios with strong departures from the perfect fit, β_2 was selected to attain $\hat{\epsilon} \approx 10\epsilon_0$ in samples of 1 million records. The upper panels in Figure 3 provide a series of scatterplots comparing a model's predictions to the true probabilities in one simulated dataset per scenario.

The second family of simulations focused on the omission of an interaction term. The design was analogous to the one of the first set of simulations. However, the true probabilities were generated using the model logit $(p_i) = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i1} X_{i2}$, with X_{i1} and X_{i2} being distributed as a standard normal and a Bernoulli with probability 0.5, respectively. The fitted model did not include the interaction term. The values chosen for the coefficients are reported in Table 1 and scatterplots comparing model's predictions and true probabilities are reported in Figure 3. Similar to the first set of simulations, the interaction coefficient, γ_3, was tuned to obtain similar levels of $\hat{\epsilon}$ (model misspecification) across outcome rates.

The third set of simulations studied the lack of fit introduced by a misspecified link function. Outcome values were sampled using probabilities defined by logit $(p_i) = h(\tau_0 + \tau_1 X_{i1}; \alpha_1, \alpha_2)$, with $h(\cdot a_1, a_2)$ being a function belonging to the family proposed by Stukel (1988). The author introduced a class of nonlinear functions indexed by two shape parameters, α_1 and α_2, to study deviations from the traditional logistic model. In particular, the choice $\alpha_1 = 0$ and $\alpha_2 = 0$ corresponds to $h\ (x; \alpha_1, \alpha_2) = x$, which implies a logit link. Values of the parameters increasingly different from zero introduce stronger departures from the logistic function. We fit the model that assumed the logit link and chose values for α_1 and α_2 to simulate models with acceptable and poor fit. As was the case in the previous scenarios, the values of the parameters are reported in Table 1, while Figure 3 compares predicted and true probabilities. As in the other two simulation scenarios, α_1 and α_2 were tuned to obtain similar levels of $\hat{\epsilon}$ across outcome rates.

For each scenario, we considered five sample sizes: n = 25 000, 50 000, 100 000, 500 000, and 1 million. For each combination of sample size and scenario, we generated 1000 datasets.

Finally, to evaluate the sensitivity of the results to the choice of the number of groups G and to the number of covariates in the models, we ran two additional families of simulations in the scenario where the quadratic term was omitted and the outcome rate was intermediate; that is, the true model was as specified in scenarios 4-6 in Table 1. First, we increased the number of groups to $G = 50$. Second, we added 10 standard normal covariates to each fitted model.

4.2 | Results

We provide the average value of $\hat{\epsilon}$ across the 27 scenarios in Web Figure 1. The average $\hat{\epsilon}$ is, overall, constant across sample sizes. The most pronounced variations are observed in the scenarios with small or no misspecification, in correspondence of the smallest sample sizes. This behavior is caused by the zero-truncated estimator of λ (see Equation(3)). If the fitted model is correct, the true value of λ is 0, but the estimates $\hat{\lambda}$ are greater or equal to 0. When the sample size is small, the variability of $\hat{\lambda}$ is large and many estimates will be positive. This causes the average value of $\hat{\epsilon}$ to be strictly greater than 0. Increasing the sample size reduces the variability of $\hat{\lambda}$and

TABLE 1 Scenarios considered in the simulation study

Misspecification	Outcome rate	Magnitude of misspecification	True model	Index
Omission of quadratic term	1%	None	logit $(p) = -5 + X_1$	1
	1%	Small	logit $(p) = -5 + X_1 + 0.025X_1^2$	2
	1%	Large	logit $(p) = -5 + X_1 + 0.15X_1^2$	3
	10%	None	logit $(p) = -2.5 + X_1$	4
	10%	Small	logit $(p) = -2.5 + X_1 + 0.01X_1^2$	5
	10%	Large	logit $(p) = -2.5 + X_1 + 0.07X_1^2$	6
	20%	None	logit $(p) = -1.6 + X_1$	7
	20%	Small	logit $(p) = -1.6 + X_1 + 0.008X_1^2$	8
	20%	Large	logit $(p) = -1.6 + X_1 + 0.055X_1^2$	9
Omission of interaction term	1%	None	logit $(p) = -5.5 + X_1 + X_2$	10
	1%	Small	logit $(p) = -5.5 + X_1 + X_2 + 0.11X_1X_2$	11
	1%	Large	logit $(p) = -5.5 + X_1 + X_2 + 0.6X_1X_2$	12
	10%	None	logit $(p) = -3.2 + X_1 + X_2$	13
	10%	Small	logit $(p) = -3.2 + X_1 + X_2 + 0.05X_1X_2$	14
	10%	Large	logit $(p) = -3.2 + X_1 + X_2 + 0.3X_1X_2$	15
	20%	None	logit $(p) = -2.2 + X_1 + X_2$	16
	20%	Small	logit $(p) = -2.2 + X_1 + X_2 + 0.038X_1X_2$	17
	20%	Large	logit $(p) = -2.2 + X_1 + X_2 + 0.25X_1X_2$	18
Misspecified link	1%	None	logit $(p) = h\,(-5 + X_1;\ \alpha_1 = \alpha_2 = 0) = -5 + X_1$	19
	1%	Small	logit $(p) = h(-5 + X_1;\ \alpha_1 = \alpha_2 = 0.05)$	20
	1%	Large	logit $(p) = h(-3.5 + X_1;\ \alpha_1 = \alpha_2 = 0.3)$	21
	10%	None	logit $(p) = h\,(-2.5 + X_1;\ \alpha_1 = \alpha_2 = 0) = -2.5 + X_1$	22
	10%	Small	logit $(p) = h(-2.5 + X_1;\ \alpha_1 = \alpha_2 = 0.02)$	23
	10%	Large	logit $(p) = h(-2.5 + X_1;\ \alpha_1 = \alpha_2 = 0.15)$	24
	20%	None	logit $(p) = h\,(-1.6 + X_1;\ \alpha_1 = \alpha_2 = 0) = -1.6 + X_1$	25
	20%	Small	logit $(p) = h(-1.6 + X_1;\ \alpha_1 = \alpha_2 = 0.01)$	26
	20%	Large	logit $(p) = h(-1.6 + X_1;\ \alpha_1 = \alpha_2 = 0.12)$	27

Note. We fit the model logit $(p) = \beta_0 + \beta_1 X_1$in scenarios 1-9 (omission of quadratic term) and 19-27 (misspecified link), while we fit the model logit $(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ in scenarios 10-18 (omission of the interaction term)

causes the average of the $\hat{\epsilon}$ to shrink toward zero. This phenomenon is represented in Web Figure 2, which reports the distributions of the estimated $\hat{\epsilon}$ at the different sample sizes for scenarios 4-6.

Figure 4 reports the number of rejections of the three methods in the scenarios with outcome rate equal to 10%. When the model is correctly specified, the rejection rate of the HL test is very close to the nominal level of 5% (solid lines in left panels). As expected, the number of rejections increases with increasing sample size for the misspecified models. In particular, despite the similarity of predicted and true probabilities in scenarios 5, 14, and 23, the average rejection rate in these scenarios is about 50% in sample sizes of 1 million observations (dotted lines in left panels). In other words, although these models would be perfectly acceptable in most applications, the HL test indicates that the models fit poorly about 50% of the time.

The resampling approach proposed by Lai and Liu perfectly controlled the type-I error rate (solid lines in right panels). Moreover, because the rejection rate is approximately constant in the misspecified models (dotted and dashed lines in right panels), this approach appears to standardize the power across sample sizes. We find low rejection rates in scenarios 5, 14, and 23, characterized by minor misspecifications of the fitted models (dotted lines in right panels). However, the power to detect important misspecifications—scenarios 6, 15 and 24—is also very low (dashed lines in right panels), and similar in magnitude to the power in scenarios with minor departures from the perfect fit. Because of the limited size of the subsamples (1,000 observations), this resampling strategy is not capable of discerning between models with acceptable and very poor fit.

Our approach overcomes the limitations of the HL test and Lai and Liu approach. First of all, the rejection rate

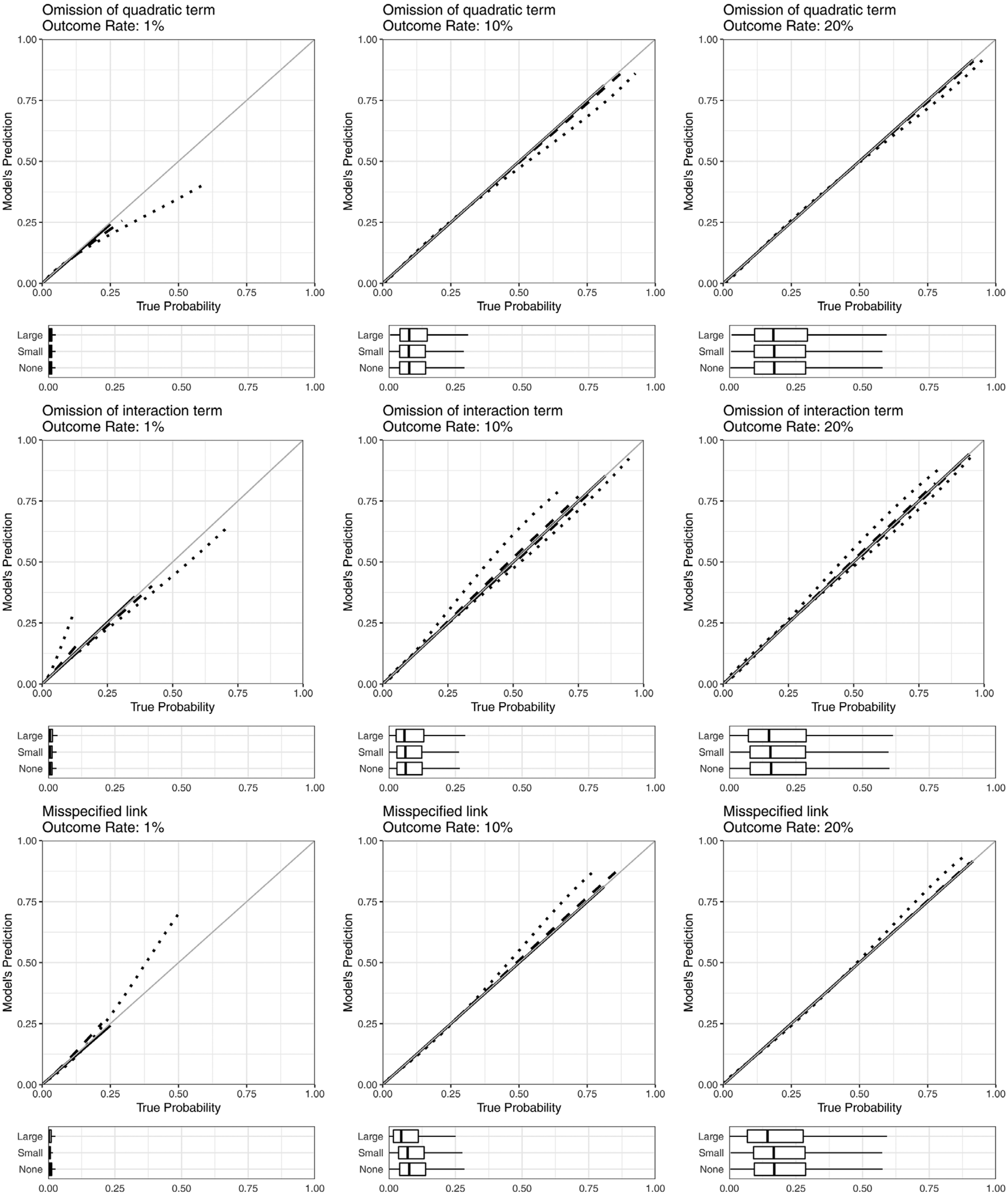


FIGURE 3 Graphical representation of the simulation scenarios. Upper, central, and bottom panels correspond to the three families of misspecifications. Left, central, and right panels correspond to the three outcome rates. Each panel compares the estimated and true probabilities in one simulated dataset for three models: correctly specified model (solid line), model with small departure from perfect fit (dashed line), and model with large departure from perfect fit (dotted line). Below each scatterplot, the boxplots summarize the distribution of the true probabilities

of scenarios 5, 14, and 23 was very small and close to 5% (dotted lines in the central panels). This result was expected. The parameters in scenarios 5, 14, and 23 were tuned to be characterized by a value of ϵ approximately equal to ϵ_0. Because our test evaluates the hypotheses $H_0 : \epsilon \leq \epsilon_0$ versus $H_a : \epsilon > \epsilon_0$, rejection rates close to the nominal level of 5% are expected when the true value of ϵ is exactly ϵ_0. In particular, if the true value of ϵ is smaller than ϵ_0, the rejection rate is expected to be smaller than the adopted significance. This is the reason for the low rejection rates of our approach in scenarios where the model is correctly specified (solid lines in central panels). In these scenarios (4, 13, and 22), the true value of ϵ is 0 and rejection rates lower than 5% are consequently expected. Importantly, the power to detect important misspecifications of the model is high and very similar to the power of the traditional HL test.

Very similar results were found in scenarios with outcome rates of 1% and 20%. For these cases, we provide outputs similar to Figure 4 in Web Figures 3 and 4. Increasing the number of groups and the number of covariates in the models did not affect the results. When G was increased to 50, the proposed approach controlled the rejections of the models characterized by small misspecifications (Web Figure 5). Very similar results were found when 10 additional predictors were considered in the models (Web Figure 6).

5 | APPLICATION: A PREDICTIVE MODEL FOR POSTNEONATAL MORTALITY

The Infant Mortality Research Partnership is a collaborative effort to reduce infant mortality in Ohio (Ohio Colleges of Medicine Government Resource Center, 2018). The project is sponsored by state agencies and includes multidisciplinary teams of researchers, who leveraged a variety of statistical methods to identify targeted, effective interventions. A family of predictive models was developed to serve as point-of-care tools for health-care providers. Logistic regression models were built to estimate the probability of poor pregnancy outcomes, such as preterm delivery or infant mortality. A rich dataset linking insurance claims and vital statistics data was made available to researchers after deidentification.

Included in the set of developed models is one that estimates the probability of postneonatal mortality based on variables available at the six-week postpartum visit. The model was developed on a sample of 315 828 babies. Twelve independent variables were selected as the model's predictors, following best practices and recommendations (Hosmer *et al.*, 2013). These variables included mother's demographics, chronic conditions, obstetric history, behaviors during pregnancy, and information about the delivery. Among the predictors, two were continuous variables: the age of the mother and the gestational age of the baby at birth.

A first version of the model included the selected predictors, with the continuous variables in linear form, and no interaction. We refer to this model as Model 1. A value of the area under the ROC curve of .758 showed good discrimination. However, a P-value of .001 in the HL test suggested poor calibration (HL statistic = 25.35, $G = 10$ groups). The model-building process was completed by assessing the linearity in the logit of continuous predictors and considering possible interactions. A quadratic term of mother's age was added to the model, because of a nonlinear, U-shaped relationship between mother's age and the logit of the probability of postneonatal mortality. The final model also included a clinically relevant interaction, between the binary variable indicating whether the baby was transferred to a different hospital at birth and his/her gestational age. We refer to the final model as Model 2. The discrimination of this model was very similar to the simpler model (area under the ROC curve: .760). On the other hand, the calibration was considerably improved (HL P-value: .034, statistic = 16.66). Nevertheless, the test still points to suboptimal goodness of fit. Because of the very large sample size and the P-value being close to the traditional significance level of 0.05, researchers might argue that the calibration of the model was within a range of acceptability.

We now evaluate the goodness of fit of the two models with the proposed methodology. First, we estimated ϵ. For Model 1, we have $\hat{c} = 25.35$, $G = 10$, and $n = 315,828$ which, when plugged into Equation (3), gives

$$\hat{\epsilon}_1 = \sqrt{\frac{\max\{25.35 - (10 - 2),\ 0\}}{315,828}} = 7.41 \times 10^{-3}. \qquad (5)$$

In a similar fashion, we estimated $\hat{\epsilon}_2 = 5.24 \times 10^{-3}$ for Model 2. The corresponding two-sided 95% CIs for ϵ_1 and $\epsilon_2$were computed using the numerical procedure described in Section 3.1. These intervals are displayed in Figure 5 along with a curve that shows the value of $\hat{\epsilon}$ for a model attaining a HL P-value of .05 across different sample sizes. Such a curve is easily obtained by varying the reference size n_0 in Equation (4). In particular, values of $\hat{\epsilon}$ above the curve correspond to models rejected at the .05-level based on the HL test. This is the case of the models under assessment (black and gray dots). A first qualitative assessment of the goodness of fit can be carried out by comparing the range of values spanned by the CIs and the curve of borderline calibration at different sample sizes. In the case of Model 2, the point estimate (black dot) is extremely close to the curve. In particular, if the same value of $\hat{\epsilon}_2$ had been obtained with a slightly smaller sample size—say, $n = 250\,000$—the dot would have been below the curve, and the HL test would have failed to reject the null hypothesis. Moreover, the CI (black error bar) covers the range of values attained by the curve for very large samples, such as $n = 10^6$. This suggests that the fit of Model 2 is not different from the fit of a model that would attain a HL P-value of .05 in a sample of 1 million observations. The qualitative assessment of Model 1 reaches different conclusions. The gray error bar spans appreciably higher values and, for example, suggests strong evidence that the fit of Model 1 is worse than a borderline model in a sample of size 500 000.

Such a visual assessment can be amended with a formal test. If we consider a model that would attain a HL P-value of .05 in a sample of size $n_0 = 10^6$ as acceptable, the reference value ϵ_0 is the one provided in Equation (4): $\epsilon_0 = 2.74 \times 10^{-3}$.

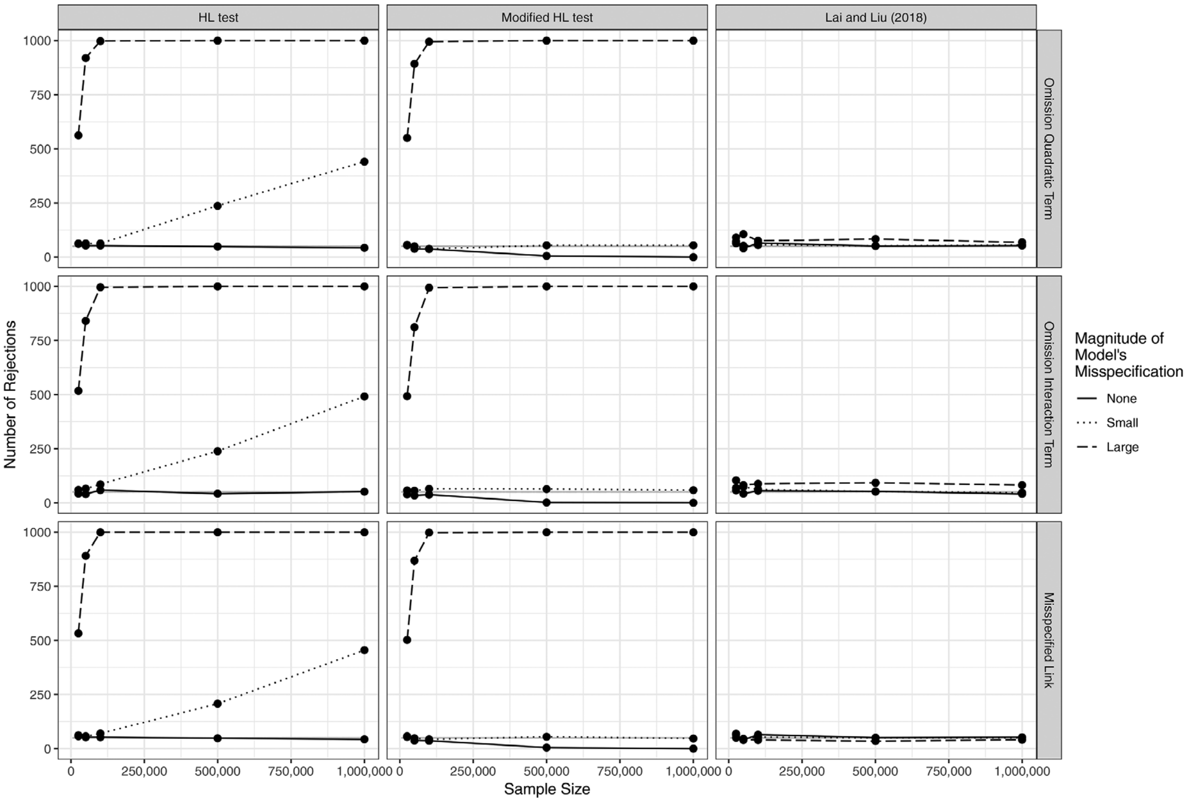


FIGURE 4 Number of rejections (out of 1000 simulated dataset) of the HL test (left panels), proposed test (central panels), and Lai and Liu resampling approach (right panels) in the scenarios with 10% outcome rate. Upper, central, and bottom panels correspond to the three families of misspecifications. Within each panel, solid, dotted, and dashed lines represent scenarios with no, small, and large model misspecification. The gray line depicts the 5% level

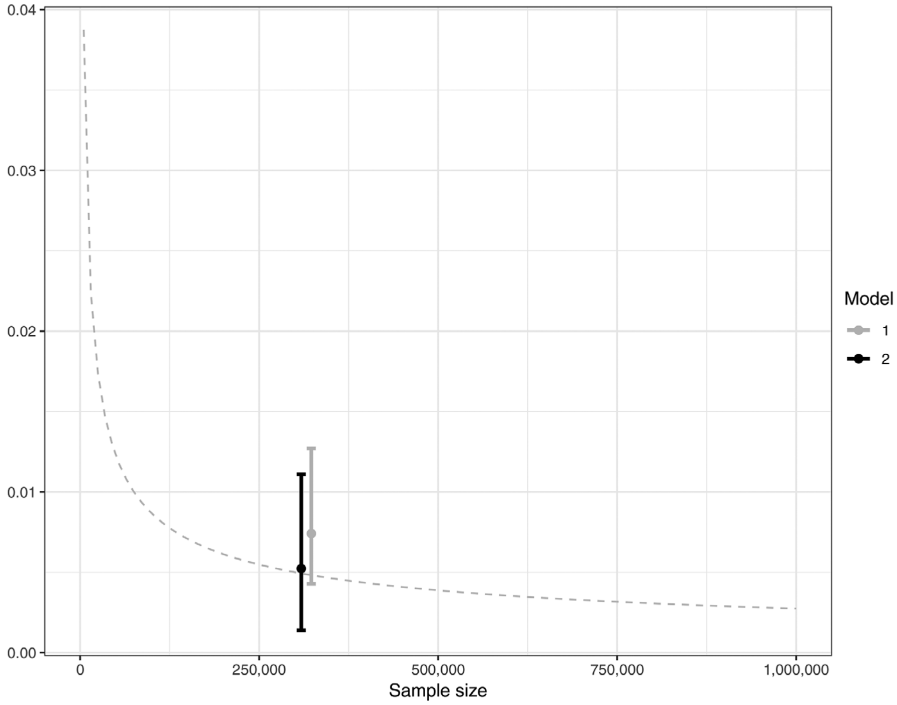


FIGURE 5 Estimates of $\hat{\epsilon}$ for the two postneonatal mortality models and corresponding two-sided 95% confidence intervals. Although the models are built on the same sample, the error bars are slightly shifted on the x-axis to avoid being superimposed. The gray dashed line represents the value attained by a model characterized by a P-value of .05 at the different sample sizes

To formally test whether the goodness of fit of the two models is acceptable—that is, testing the hypothesis $H_0 : \epsilon \leq \epsilon_0$ versus $H_a : \epsilon > \epsilon_0$—we need to compute the one-sided CIs of ϵ. The lower bounds of these CIs can be computed numerically and are 4.28×10^{-3} and 1.39×10^{-3} for models 1 and 2, respectively. Because ϵ_0 is contained in the second interval, there is insufficient evidence (at the 0.05-level) that the fit of Model 2 is not acceptable. Conversely, the CI corresponding to Model 1 does not contain the value ϵ_0, which suggests strong evidence of a non-acceptable fit.

Using the result in Proposition 2, we can also compute a P-value associated with this test, which can be regarded as the formal modification of the HL test. This P-value is computed by comparing the value of the traditional HL statistic with a noncentral chi-squared distribution with $G - 2$ degrees of freedom (as in the traditional test) and noncentrality parameter $\epsilon_0^2 n$, which is equal to $(2.74 \times 10^{-3})^2 \times 315\,828 = 2.37$ in our case. The P-values of the modified HL test for models 1 and 2 are .010 and .11, respectively.

These results confirm the heuristic intuition of the researchers who developed the model with a rigorous testing procedure. The proposed modification of the HL test suggests that there is strong evidence of poor fit of the simpler model without nonlinear terms and interactions (Model 1). Conversely, we can conclude that there is insufficient evidence that the fit of the final model (Model 2) is not acceptable for practical purposes.

6 | DISCUSSION

Because the power of goodness of fit testing procedures, such as the HL test, increases with the sample size, minor model misspecifications may cause the rejection of the hypothesis of perfect fit in very large samples. We propose a modification of the HL test that allows a rigorous evaluation of fit unaffected by sample size. Such a modification is based on a parameter, ϵ, which standardizes the noncentrality parameter of the HL statistic with respect to the sample size and can be used to measure the lack of fit of the model. The proposed methodology is appropriate to entirely replace the HL test, for any sample size, as the two tests rely on the same assumptions. We show that the results of the two tests are practically identical for small-to-moderate samples, whereas they differ substantially only if the sample size is very large.

If the values of ϵ had a meaningful interpretation, the estimation of this parameter would be sufficient to characterize the goodness of fit of a model. By comparing the estimate and the CI of ϵ with the range of acceptable values of the parameter, we would be able to make decisions about the model's fit. Unfortunately, the values of ϵ do not have immediate interpretation. We propose a simple method to identify a value ϵ_0 corresponding to an acceptable, albeit not perfect, model, that is, a model that would correspond to a P-value of the HL test of .05 in a sample of size 1 million. Although we feel that our choice is reasonable, it might not suit all the areas of application. For example, the reference size n_0 of 1 million might not be large enough in studies requiring very well calibrated models, such as in large epidemiological studies using the standardized mortality ratio to compare the performance of hospitals (Kipnis *et al.*, 2014). On the other hand, the choice of 1 million might be too restrictive in contexts where the discrimination of the model is of primary importance, such as in studies developing diagnostic tests. To assess the sensitivity of the methodology to the choice of n_0, one may compute P-values using the HL statistic computed for a particular model and a range of reference sample sizes n_0. For example, in Web Figure 7, we show how the P-value of our approach does not change qualitatively for values of n_0 spanning from 500 000 to 5 million in a sample of size $n = 100\,000$.

Our simulations show that the proposed test rejects less than the nominal level when the correct model is evaluated. Such conservative behavior would be worrisome only if the power under an important model misspecification were low. On the contrary, the power of the proposed test is similar to the power of the traditional approach in scenarios with important departures from perfect fit. In these same scenarios, the power of the resampling approach of Lai and Liu (2018) was very low. This result is likely due to the small size of the considered subsamples and raises doubts about the usefulness of resampling methods to detect important model misspecifications.

We focused on the assessment of logistic regression models developed on independent, equally weighted observations. However, logistic regression is also used in cases where the data come from complex survey designs or where observations have unequal weights. Archer *et al.* (2007) proposed a generalization of the HL test for this setting. Because survey studies often result in very large samples, extensions of our methodology to complex survey data should be promptly investigated.

ORCID

Giovanni Nattino https://orcid.org/0000-0002-3034-6251

REFERENCES

Archer, K.J., Lemeshow, S. and Hosmer, D.W. (2007) Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics & Data Analysis*, 51(9), 4450–4464.

Austin, P.C. and Steyerberg, E.W. (2014) Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*, 33(3), 517–535.

Browne, M.W. and Cudeck, R. (1992) Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230–258.

Casella, G. and Berger, R.L. (2002) *Statistical Inference*. Pacific Grove, CA: Thomson Learning.

Dahiya, R.C. and Gurland, J. (1973) How many classes in the Pearson chi-square test? *Journal of the American Statistical Association*, 68(343), 707–712.

Hosmer, D.W. and Lemesbow, S. (1980) Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics – Theory and Methods*, 9(10), 1043–1069.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) *Applied Logistic Regression*. Chicester, UK: Wiley.

Kent, J.T. and Hainsworth, T.J. (1995) Confidence intervals for the noncentral chi-squared distribution. *Journal of Statistical Planning and Inference*, 46(2), 147–159.

Kipnis, P., Liu, V. and Escobar, G.J. (2014) Accuracy of hospital standardized mortality rates: effects of model calibration. *Medical Care*, 52(4), 378–384.

Kramer, A.A. and Zimmerman, J.E. (2007) Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. *Critical Care Medicine*, 35(9), 2052–2056.

Lai, X. and Liu, L. (2018) A simple test procedure in standardizing the power of Hosmer-Lemeshow test in large data sets. *Journal of Statistical Computation & Simulation*, 88(13), 2463–2472.

Moons, K.G.M., Altman, D.G., Reitsma, J.B., Ioannidis, J.P.A., Macaskill, P., Steyerberg, E.W., Vickers, A.J., Ransohoff, D.F. and Collins, G.S. (2015) Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine*, 162(1), W1–W73.

Moore, D.S. and Spruill, M.C. (1975) Unified large-sample theory of general chi-squared statistics for tests of fit. *The Annals of Statistics*, 3(3), 599–616.

Nattino, G., Finazzi, S. and Bertolini, G. (2016) A new test and graphical tool to assess the goodness of fit of logistic regression models. *Statistics in Medicine*, 35(5), 709–720.

Ohio Colleges of Medicine Government Resource Center. (2018) *Infant mortality research partnership*. Available at: https://grc.osu.edu/projects/IMRP [Accessed 15 November 2018].

Paul, P., Pennell, M.L. and Lemeshow, S. (2013) Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Statistics in Medicine*, 32(1), 67–80.

Saxena, K.M.L. and Alam, K. (1982) Estimation of the non-centrality parameter of a chi squared distribution. *The Annals of Statistics*, 10(3), 1012–1016.

Steiger, J.H. (1990) Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research*, 25(2), 173–180.

Steiger, J.H. and Fouladi, R.T. (1997) Noncentrality interval estimation and the evaluation of statistical models. In: Harlow, L.L., Mulaik, S.A. and Steiger, J.H. (Eds). *What if There were no Significance Tests*. Abingdon, UK: Taylor & Francis, pp. 197–229.

Stukel, T.A. (1988) Generalized logistic models. *Journal of the American Statistical Association*, 83(402), 426–431.

Yu, W., Xu, W. and Zhu, L. (2017) A modified Hosmer–Lemeshow test for large data sets. *Communications in Statistics – Theory and Methods*, 46(23), 11813–11825.

SUPPORTING INFORMATION

Web Appendices and Figures referenced in Sections 2, 3.2, 4.2 and 6 and the code generating the results of the paper are available with this paper at the Biometrics website on Wiley Online Library.

The R package largesamplehl that implements the methodology is available at the Github repository https://github.com/gnattino/largesamplehl.

How to cite this article: Nattino G, Pennell ML, Lemeshow S. Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test. *Biometrics*. 2020;76:549–560. https://doi.org/10.1111/biom.13249