The New England Journal of Statistics in Data Science Volume 0, 1–15 (2023)
DOI: https://doi.org/10.51387/23-NEJSDS56

# A Safe Hosmer-Lemeshow Test

Alexander HENZI[1], Marius PUKE[1,*], Timo DIMITRIADIS, and Johanna ZIEGEL

**Abstract**

This article proposes an alternative to the Hosmer-Lemeshow (HL) test for evaluating the calibration of probability forecasts for binary events. The approach is based on e-values, a new tool for hypothesis testing. An e-value is a random variable with expected value less or equal to one under a null hypothesis. Large e-values give evidence against the null hypothesis, and the multiplicative inverse of an e-value is a p-value. Our test uses online isotonic regression to estimate the calibration curve as a 'betting strategy' against the null hypothesis. We show that the test has power against essentially all alternatives, which makes it theoretically superior to the HL test and at the same time resolves the well-known instability problem of the latter. A simulation study shows that a feasible version of the proposed eHL test can detect slight miscalibrations in practically relevant sample sizes, but trades its universal validity and power guarantees against a reduced empirical power compared to the HL test in a classical simulation setup. We illustrate our test on recalibrated predictions for credit card defaults during the Taiwan credit card crisis, where the classical HL test delivers equivocal results.

keywords and phrases: E-value, Probability forecast, Calibration validation, Goodness-of-fit, Isotonic regression.

## 1. INTRODUCTION

Suppose that we have observations $(p_i, y_i)_{i=1}^n$ of independent and identically distributed (iid) random variables $(P_i, Y_i)_{i=1}^n$ with $(P_i, Y_i) \in [0,1] \times \{0,1\}$, $i = 1, \dots, n$. The interpretation is that $P_i$ is a prediction for the probability that $Y_i = 1$. The random variables are defined on some underlying probability space $(\Omega, \mathcal{F})$ and $\mathcal{P}$ denotes all probability measures on $(\Omega, \mathcal{F})$. Hosmer and Lemeshow [18] propose a test for the null hypothesis of perfect calibration

$$\mathcal{H}_{\mathrm{HL},n} = \{\mathbb{P} \in \mathcal{P} \mid \mathbb{E}_{\mathbb{P}}(Y_i|P_i) = P_i \quad \mathbb{P}\text{-almost surely, } i = 1, \dots, n\}. \tag{1.1}$$

The Hosmer-Lemeshow (henceforth HL) test is based on partitioning the interval $[0,1]$ in $g \in \mathbb{N}$ bins and counting the observed numbers of events, $o_{1g}$, and no event occurrences, $o_{0g}$, in each bin. Based on that binning and counting procedure, the HL test statistic to test for perfect calibration of the probability predictions is

$$\widehat{C} = \sum_{k=1}^{g} \left[ \frac{(o_{1k} - \widehat{e}_{1k})^2}{\widehat{e}_{1k}} + \frac{(o_{0k} - \widehat{e}_{0k})^2}{\widehat{e}_{0k}} \right], \tag{1.2}$$

where $\widehat{e}_{1k}$ and $\widehat{e}_{0k}$ are the expected event and no event occurrences in bin $k$, respectively [19]. Under the null hypothesis, $\widehat{C}$ asymptotically follows a $\chi^2$-distribution with $g$ degrees of freedom given that the sample $(P_i, Y_i)_{i=1}^n$ was not used for model estimation (and $g-2$ degrees of freedom otherwise).

Technically, the choice of the binning procedure is up the user of the HL test and is conventionally implemented via quantile based binning strategies with $g = 10$, resulting in equally populated bins (decile-of-risk). Less commonly, the test is based on equidistantly spaced bins, where the unit interval (or the range of prediction values) is divided into $g$ equidistant bins. While little attention is devoted to the binning procedure in practical applications, it implicitly determines the set of alternatives the test has power against [8, Section 5], such that the test result is often highly sensitive to the exact implementation of the binning; see e.g., [17, 3, 22] and our empirical application in Section 4. Nevertheless, the HL test is still the literature's favorite for checking the calibration of binary prediction models and commonly used in current and highly influential medical and epidemiological studies; see amongst many others [26, 28, 23].

In this article, we suggest a safe and stable HL test based on e-values (that we describe below) and isotonic regression [2, 5]. The test is henceforth called eHL test. Dimitriadis et al. [9] recently propose the use of isotonic regression to resolve the closely related instability issue stemming from binning approaches in so-called reliability diagrams in forecast evaluation. While feasible inference on the isotonic regression for classical testing procedures is hampered by complicated asymptotic distributions and an inconsistency of the bootstrap, the e-values adopted here prove to be an appealing alternative in this setting. Based on online isotonic regression studied by [20], we show that (an ideal version of) our eHL test has power against essentially all deviations from calibration, which makes it theoretically superior to the classical HL test.

*Corresponding author.
[1]The first two authors contributed equally to this work.

E-values, where 'e' abbreviates the word 'expectation', were proposed recently as an alternative to p-values in testing problems. In a nutshell, an e-value is a realization of a non-negative random variable whose expected value is at most one under a given null hypothesis. This already signals that an e-value itself allows for meaningful interpretations since an e-value greater than one provides evidence against the null hypothesis. Additionally, the multiplicative inverse of an e-value is a conservative p-value by Markov's inequality. From a game-theoretic perspective, the e-value has a simple financial meaning in the sense that the e-value can be seen as the factor by which a skeptic multiplies her money when betting against the null hypothesis; see [32, 31].

An important advantage of e-values over p-values is their uncomplicated behavior in combinations: the arithmetic average of e-values also is an e-value, likewise the product of independent or successive e-values; see [31, 13, 39]. In practice, this appeals because more evidence can be added later, i.e. evidence across studies can easily be combined.

The proposed eHL test offers a safe alternative to a fragile state-of-the-art approach by avoiding ad-hoc choices and software instabilities. It can be regarded as an application of the Universal Inference approach of [40]. While this method allows to construct valid tests under only weak assumptions, it has been observed that this validity often comes at the price of a diminished power [34, 35]. In Section 2, we show that an ideal – but computationally infeasible – variant of the eHL test does have guaranteed power to detect essentially all violations of calibration. Our proof relies on connections between the proposed e-value and the regret in random permutation online isotonic regression, which is studied by [20]. It has been observed that power guarantees for anytime-valid tests can be obtained by means of regret bounds of online prediction methods, see for example the discussion in [7]. Previously, [27] and [33] exploit that connection in the cases of sequential mean and two sample testing, respectively. Our result demonstrates that such a connection also exists in the batch case of $e$-values for a fixed sample size $n$ due to connections with the online random permutation setting.

In Section 3, we compare a feasible version of the eHL test to the classical HL test in a simulation study. As expected, we find that the eHL test has conservative rejection rates under the null hypothesis and quickly develops power under model misspecification. While its empirical test power is lower than the one of the classical HL test, we do not consider this to be problematic as HL tests are often carried out in cases of vast data sets and are even criticized as being "too powerful" in that they reject essentially all, even acceptably well calibrated models [29, 25]. See [8] for an alternative solution to this problem based on confidence bands.

We apply the eHL test in Section 4 to predictions of a logistic regression model for the binary event of credit card defaults in Taiwan in 2005, where over-issuing of credit cards lead to many default payments and a subsequent credit card crisis [43]. The eHL test provides clear evidence against calibration of the logistic model predictions, and further illustrates that recalibration methods work well. In contrast, the classical HL test based on different natural binning choices delivers equivocal results with p-values ranging from 0 to 0.91 for a single prediction method, implying that a researcher could have cherry-picked the binning specification and hence the test result to her will.

## 2. CONSTRUCTION OF HL E-VALUES

### 2.1 Preliminaries

An e-variable for $\mathcal{H}_{\mathrm{HL},n}$ is a non-negative random variable $E$ (that is allowed to take the value $+\infty$) such that $\mathbb{E}_{\mathbb{P}}(E) \leq 1$ for all $\mathbb{P} \in \mathcal{H}_{\mathrm{HL},n}$. An e-value is a realization of an e-variable. An e-variable $E$ always yields a valid p-variable $1/E$ (a p-value is a realized p-variable) by Markov's inequality, since

$$\mathbb{P}\Big(\frac{1}{E} \leq \alpha\Big) \leq \alpha \mathbb{E}_{\mathbb{P}}(E) \leq \alpha, \;\; \mathbb{P} \in \mathcal{H}_{\mathrm{HL},n}. \tag{2.1}$$

We reject the null hypothesis $\mathcal{H}_{\mathrm{HL},n}$ if we observe a large value of $E$. If we want to ensure a classical p-guarantee then we have to determine the rejection region for a given $\alpha$ by (2.1). Vovk and Wang [38] show that this is essentially the only way to transform an e-variable into a p-variable. We say that an e-variable has the alternative hypothesis $\mathcal{H}' \subset \mathcal{P}$ if $\mathbb{E}_{\mathbb{Q}}(E) > 1$ for all $\mathbb{Q} \in \mathcal{H}'$.

### 2.2 Sample Size One

We first construct e-variables for the sample size one Hosmer-Lemeshow null hypothesis

$$\mathcal{H}_{\mathrm{HL},1} = \{\mathbb{P} \in \mathcal{P} \mid \mathbb{E}_{\mathbb{P}}(Y|P) = P\}.$$

In the special case here, e-variables are likelihood ratios conditional on $P$. Indeed, if $q \in [0,1]$, an e-variable for $\mathcal{H}_{\mathrm{HL},1}$ is given by

$$E_q(P,Y) = \frac{q^Y(1-q)^{1-Y}}{P^Y(1-P)^{1-Y}} = \begin{cases} q/P, & \text{if } Y = 1, \\ (1-q)/(1-P), & \text{if } Y = 0. \end{cases}$$

The variable $E_q(P,Y)$ is clearly non-negative, and for $\mathbb{P} \in \mathcal{H}_{\mathrm{HL},1}$,

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}(E_q(P,Y)) &= \mathbb{E}_{\mathbb{P}}\left(\mathbb{E}_{\mathbb{P}}(Y \mid P)\frac{q}{P} + \mathbb{E}_{\mathbb{P}}(1-Y \mid P)\frac{1-q}{1-P}\right) \\ &= \mathbb{E}_{\mathbb{P}}\left(P\frac{q}{P} + (1-P)\frac{1-q}{1-P}\right) = 1. \end{aligned}$$

To find alternative hypotheses for the e-variable $E_q$, let $\bar{\pi} = \mathbb{E}_{\mathbb{Q}}(Y \mid P)$. Then,

$$\mathbb{E}_{\mathbb{Q}}(E_q(P,Y) \mid P) = \bar{\pi}\frac{q}{P} + (1-\bar{\pi})\frac{1-q}{1-P}$$

is strictly larger one if and only if, $\bar{\pi} > P$ and $q > P$, or, $\bar{\pi} < P$ and $q < P$, i.e., if $\pi$ and $q$ are to the same side of $P$. This shows that if $q < P$, $E_q$ has the alternative

$$\mathcal{H}' = \{\mathbb{Q} \in \mathcal{P} \mid \mathbb{E}_{\mathbb{Q}}(Y \mid P) < P\}, \tag{2.2}$$

and if $q > P$, $E_q$ has the alternative

$$\mathcal{H}' = \{\mathbb{Q} \in \mathcal{P} \mid \mathbb{E}_{\mathbb{Q}}(Y \mid P) > P\}. \tag{2.3}$$

It is possible to show that basically any e-variable for $\mathcal{H}_{\mathrm{HL},1}$ is of the form $E = E_q(P, Y)$ for some $q$ (depending on $P$) but this requires some more arguments; it follows by the construction in [15], see also [41]. The connection of $E_q(P, Y)$ to the e-variables in [15] of type $E = 1 + \lambda D$ with $D \geq -1$ such that $\mathbb{E}_{\mathbb{P}}(D) = 0$ for $\mathbb{P} \in \mathcal{H}_{\mathrm{HL},1}$, follows from the fact that $\lambda$ in this representation can be bijectively mapped to $q$. In this context,

$$E = 1 + \lambda(P - Y) \tag{2.4}$$

is an e-variable for $\mathcal{H}_{\mathrm{HL},1}$ for any $\lambda$ that is $\sigma(P)$-measurable with $-(1/P) \leq \lambda \leq 1/(1-P)$. If $P = 1$, there is no restriction on $\lambda$ from above, and analogously if $P = 0$, there is no restriction from below. By choosing $\lambda = (P-q)/(P(1-P))$, we obtain that $E = E_q(P, Y)$.

Clearly, the e-variable $E_q(P, Y)$ may take the value infinity if either $P = 0$ and $Y = 1$ or $P = 1$ and $Y = 0$ occurs; a single observation $Y = 1$ or $Y = 0$ is sufficient to reject the hypothesis of calibration with certainty if the predicted probabilities are in $\{0, 1\}$. For the remainder of the theoretical part of this paper, we will always make the assumption $\mathbb{P}(P \in \{0, 1\}) = 0$ to exclude these special but uninteresting cases.

## 2.3 Combining e-Values in the iid Case

For testing $\mathcal{H}_{\mathrm{HL},n}$, we suggest the e-variable

$$E^{\mathrm{id}}_{\mathrm{HL},n} = \prod_{i=1}^{n} E_{q_i}(P_i, Y_i), \tag{2.5}$$

where $q_i$ is $\sigma(P_1, \ldots, P_i, Y_1, \ldots, Y_{i-1})$-measurable. For $\mathbb{P} \in \mathcal{H}_{\mathrm{HL},n}$, the expectation $\mathbb{E}_{\mathbb{P}} E^{\mathrm{id}}_{\mathrm{HL},n}$ equals

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}\Big(\mathbb{E}_{\mathbb{P}}\Big(\prod_{i=1}^{n} E_{q_i}(P_i, Y_i) | P_1, \ldots, P_n, Y_1, \ldots, Y_{n-1}\Big)\Big) \\
&= \mathbb{E}_{\mathbb{P}}\Big(\prod_{i=1}^{n-1} E_{q_i}(P_i, Y_i) \\
&\quad \times \mathbb{E}_{\mathbb{P}}\Big(E_{q_n}(P_n, Y_n) | P_1, \ldots, P_n, Y_1, \ldots, Y_{n-1}\Big)\Big) \\
&= \mathbb{E}_{\mathbb{P}}\Big(\prod_{i=1}^{n-1} E_{q_i}(P_i, Y_i)\Big(1 + \\
&\quad \frac{P_n - q_n}{P_n(1-P_n)} \mathbb{E}_{\mathbb{P}}(P_n - Y_n | P_1, \ldots, P_n, Y_1, \ldots, Y_{n-1})\Big)\Big) \\
&= \mathbb{E}_{\mathbb{P}}\Big(\prod_{i=1}^{n-1} E_{q_i}(P_i, Y_i)\Big(1 + \frac{P_n - q_n}{P_n(1-P_n)} \mathbb{E}_{\mathbb{P}}(P_n - Y_n | P_n)\Big)\Big) \\
&= \mathbb{E}_{\mathbb{P}}\Big(\prod_{i=1}^{n-1} E_{q_i}(P_i, Y_i)\Big) = \mathbb{E}_{\mathbb{P}} E^{\mathrm{id}}_{\mathrm{HL},n-1} = \cdots = 1,
\end{aligned}$$

where we used the equivalent representation of $E_q(P, Y)$ in (2.4). In particular, from the above derivation it is easy to see that $(E^{\mathrm{id}}_{\mathrm{HL},n})_{n \in \mathbb{N}}$ is a test martingale.

The e-variable $E^{\mathrm{id}}_{\mathrm{HL},n}$ depends on the ordering of $(P_i, Y_i)_{i=1}^n$ through the choice of $q_i$. Let $S_n$ denote all permutations of $\{1, \ldots, n\}$, and for $\sigma \in S_n$ define $E^{\sigma}_{\mathrm{HL},n}$ as $E^{\mathrm{id}}_{\mathrm{HL},n}$ for the random variables $(P_{\sigma(i)}, Y_{\sigma(i)})_{i=1}^n$ instead of $(P_i, Y_i)_{i=1}^n$. Generally,

$$\sup_{\sigma \in S_n} E^{\sigma}_{\mathrm{HL},n}$$

is not an e-variable for $\mathcal{H}_{\mathrm{HL},n}$, so one would guess that there are opportunities to fish for (spurious) significance by choosing some specific ordering of a sample of observations $(p_i, y_i)_{i=1}^n$. In constructing a 'safe' HL test, we are particularly focused on avoiding such instabilities through order-dependencies. Nevertheless, order-dependent strategies might be considered if they preclude possibilities to fish for significance.

In contrast, if there is a natural ordering of the observations such as a time stamp then the problem usually does not occur in applications since a different ordering of the observations is hard to justify. Indeed, when the observations are sequential (and possibly dependent), the e-variable defined at (2.5) is also an e-variable for the hypothesis

$$\begin{aligned}
\mathcal{H}_{\mathrm{HL},n,seq} = \{\mathbb{P} \in \mathcal{P} \mid \mathbb{E}_{\mathbb{P}}(Y_i | P_1, \ldots, P_i, Y_1, \ldots, Y_{i-1}) = P_i \\
\mathbb{P}\text{-almost surely, } i = 1, \ldots, n\}.
\end{aligned}$$

Contrary to classical theory, the sequential case is easier to treat than the iid case and has been the focus of many works employing e-values including for example [41, 15].

Coming back to our situation with iid data, an alternative to (2.5) could be

$$E_{\mathrm{HL},n,\mathrm{sym}} = \frac{1}{n!} \sum_{\sigma \in S_n} E^{\sigma}_{\mathrm{HL},n}.$$

This strategy is essentially the merging technique for independent e-values in Section 4 of [38], and the object of interest in this article.

## 2.4 An Ideal Test with Power Guarantees

The statistic $E_{\mathrm{HL},n,\mathrm{sym}}$ is an e-variable solely under the requirement that for $i = 1, \ldots, n$ and all permutations $\sigma$, the probabilities $q_{\sigma(i)}$ in $E^{\sigma}_{\mathrm{HL},n}$ are a measurable function of $(P_{\sigma(j)}, Y_{\sigma(j)})$, $j = 1, \ldots, i-1$, and of $P_{\sigma(i)}$. In the following, we write

$$q_{\sigma,\sigma(i)} = f_i(P_{\sigma(1)}, \ldots, P_{\sigma(i)}, Y_{\sigma(1)}, \ldots, Y_{\sigma(i-1)}),$$

using the same algorithm $f_i$ for constructing $q_{\sigma,\sigma(i)}$ based on $P_{\sigma(1)},\dots,P_{\sigma(i)},Y_{\sigma(1)},\dots,Y_{\sigma(i-1)}$ for all permutations $\sigma$. The challenge is then how to choose the functions $f_1,\dots,f_n$ such that the test has power. As argued by [13, 31], a suitable measure of power for e-values is the growth rate $\mathbb{E}_{\mathbb{Q}}[\log(E)]$ under an alternative distribution $\mathbb{Q}$, so that ideally, $E$ grows exponentially fast in the sample size if the null hypothesis is violated.

Our algorithm for choosing $f_1,\dots,f_n$ is inspired by permutation online isotonic regression, studied extensively by [20]. In machine learning applications, isotonic regression is an established method for the recalibration of binary classifiers; see e.g. [44] or [12]. Recently, [9] related the isotonic regression approach to reliability diagrams, which are a key diagnostic tool in evaluating probability forecast for binary events, especially in meteorology. Our results demonstrate that isotonic regression is also suitable for constructing universal tests of calibration.

To introduce the algorithm for constructing our e-variable, let $p_1,\dots,p_i\in[0,1]$ be probability predictions and $y_1,\dots,y_i\in\{0,1\}$ be the corresponding outcomes. Then the isotonic regression of $y_1,\dots,y_i$ on $p_1,\dots,p_i$ can be described as the maximizer of

$$\begin{aligned}\hat{R}_n(g_1,\dots,g_i)&=\hat{R}(g_1,\dots,g_i;\,p_1,\dots,p_i,y_1,\dots,y_i)\\&=\sum_{j=1}^{i}\log\left(\left(\frac{g_j}{p_j}\right)^{y_j}\left(\frac{1-g_j}{1-p_j}\right)^{1-y_j}\right),\end{aligned}\tag{2.6}$$

over all $g_1,\dots,g_i$ such that $g_k\le g_l$ if $p_k\le p_l$. Notice that the quantity in (2.6) is simply a normalized version of the logarithmic score, and the maximizer does not depend on the fact that we normalize by $p_j^{y_j}(1-p_j)^{1-y_j}$. Moreover notice that up to rescaling by $1/i$, this criterion also equals the sample version of $\mathbb{E}_{\mathbb{Q}}[\log(E)]$ when the e-variable $E$ is the likelihood ratio between the probabilities $g_j$ and $p_j$. A unique maximizer exists — unique since we exclude the cases $p_j=0$ and $y_j=1$ or $p_j=1$ and $y_j=0$ for some $j$ — and can be computed efficiently with the PAV-Algorithm [2]. This estimator only defines a recalibrated version of $p_1,\dots,p_i$, and a method is required to define the regression at a $p_{i+1}\in[0,1]$ not contained in the sample. To obtain out-of-sample predictions with small regret in terms of log-loss, we rely on a strategy of [30] that was adapted to isotonic regression by [37], and applied by [20] to derive regret bounds for isotonic regression in an online setting. The out-of-sample value at $p_{j+1}$ is defined as follows,

$$f_{i+1}(p_1,\dots,p_i,p_{i+1},y_1,\dots,y_i)=\frac{g_{i+1,1}}{g_{i+1,1}+1-g_{i+1,0}},\tag{2.7}$$

where $g_{i+1,1}$ and $g_{i+1,0}$ are the $(i+1)$-th component the isotonic regression of $p_i,\dots,p_i,p_{i+1}$ with observations $y_1,\dots,y_i,1$ or $y_1,\dots,y_i,0$, respectively. That is, to define the isotonic regression at the unseen $p_{i+1}$, we fit two isotonic regression in which we include $p_{i+1}$ in the sample with artificial observations of 1 and of 0 respectively, and take the ratio (2.7) as recalibrated probability. The definition (2.7) is extended to the case $i=0$ by setting $g_{1,1}=g_{1,0}=0.5$. The workflow to construct $E_{\mathrm{HL},n,\mathrm{sym}}$ is then described in Algorithm 1.

**Algorithm 1** Construction of $E_{\mathrm{HL},n,\mathrm{sym}}$.

$E_{\mathrm{HL},n,\mathrm{sym}}\leftarrow 0$
**for** all permutations $\sigma$ of $\{1,\dots,n\}$ **do**
  $E^{\sigma}_{\mathrm{HL},n}\leftarrow(0.5/P_{\sigma(1)})^{Y_{\sigma(1)}}(0.5/(1-P_{\sigma(1)}))^{1-Y_{\sigma(1)}}$
  **for** $i=1,\dots,n-1$ **do**
    $q_{\sigma,\sigma(i+1)}\leftarrow f_i(P_{\sigma(1)},\dots,P_{\sigma(i+1)},Y_{\sigma(1)},\dots,Y_{\sigma(i)})$ as defined in (2.7)
    $E^{\sigma}_{\mathrm{HL},n}\leftarrow E^{\sigma}_{\mathrm{HL},n}\cdot(q_{\sigma,\sigma,\sigma(i+1)}/P_{\sigma(i+1)})^{Y_{\sigma(i+1)}}((1-q_{\sigma,\sigma(i+1)})/(1-P_{\sigma(i+1)}))^{1-Y_{\sigma(i+1)}}$
  **end for**
  $E_{\mathrm{HL},n,\mathrm{sym}}\leftarrow E_{\mathrm{HL},n,\mathrm{sym}}+E^{\sigma}_{\mathrm{HL},n}/n!$
**end for**
**return** $E_{\mathrm{HL},n,\mathrm{sym}}$

To state a result about the power of $E_{\mathrm{HL},n,\mathrm{sym}}$, we need a population version of the isotonic regression estimator. For a function $\pi\colon[0,1]\to[0,1]$, let

$$R_{\mathbb{Q}}(\pi)=\mathbb{E}_{\mathbb{Q}}\left[\log\left((\pi(P)/P)^{Y}((1-\pi(P))/(1-P))^{1-Y}\right)\right]$$

if this expectation exists. Let $\mathcal{F}_{\uparrow,[0,1]}$ be the set of nondecreasing functions $\pi\colon[0,1]\to[0,1]$. If $\mathbb{Q}$ is the empirical distribution of $(P_1,Y_1),\dots,(P_n,Y_n)$, then it is easy to see that $R_{\mathbb{Q}}$ coincides with the target function $\hat{R}$ of the usual isotonic regression in finite samples. With these definitions, we can state the following result about the power of $E_{\mathrm{HL},n,\mathrm{sym}}$.

**Theorem 2.1.** *Let $(P_1,Y_1),\dots,(P_n,Y_n),(P,Y)$ be iid with distribution $\mathbb{Q}$ such that*

$$\mathbb{E}_{\mathbb{Q}}[\log(P)^2+\log(1-P)^2]<\infty.\tag{2.8}$$

*Then,*

*(i) there exists a $\mathbb{Q}$-almost-surely unique maximizer $\pi^*\in\mathcal{F}_{\uparrow,[0,1]}$ of $R_{\mathbb{Q}}$;*
*(ii) for a version of $\pi^*$ from part (i), let*

$$\begin{aligned}D(\mathbb{Q})&=R_{\mathbb{Q}}(\pi^*)\\&=\mathbb{E}_{\mathbb{Q}}[\log(\pi^*(P)/P)^{Y}((1-\pi^*(P))/(1-P))^{1-Y}];\end{aligned}$$

*then $D(\mathbb{Q})\ge 0$, with equality if and only if $\mathbb{Q}\in\mathcal{H}_{\mathrm{HL},1}$;*
*(iii) the e-value $E_{\mathrm{HL},n,\mathrm{sym}}$ from Algorithm 1 satisfies*

$$\begin{aligned}&E_{\mathrm{HL},n,\mathrm{sym}}\\&\quad\ge\exp\left(\sum_{i=1}^{n}\log\left(\frac{\pi^*(P_i)}{P_i}\right)^{Y_i}\left(\frac{1-\pi^*(P_i)}{1-P_i}\right)^{1-Y_i}\right.\\&\qquad\left.-C\sqrt{n\log(n)^2}\right)\end{aligned}$$

*for an universal constant $C > 0$, and hence*

$$\mathbb{E}_{\mathbb{Q}}[\log(E_{\mathrm{HL},n,\mathrm{sym}})] \geq nD(\mathbb{Q}) - C\sqrt{n\log(n)^2}.$$

The integrability assumption (2.8) is solely required to prove parts (i) and (ii) of the theorem, and the lower bound on $E_{\mathrm{HL},n,\mathrm{sym}}$ and the expectation of its logarithm in fact hold for any $\pi \in \mathcal{F}_{\uparrow,[0,1]}$. However, part (iii) only becomes useful in conjunction with (i) and (ii): the fact that $D(\mathbb{Q}) \geq 0$ with equality if and only if $\mathbb{Q} \in \mathcal{H}_{\mathrm{HL},1}$ implies that the test has positive growth rate for all alternative distributions $\mathbb{Q}$ if $n$ is large enough. This is a surprising result, since it might seem that restricting our estimator of $\mathbb{E}_{\mathbb{Q}}[Y|P]$ to isotonic functions in $P$ implies some restriction on the class of alternatives against which the test has power — which is not the case. If $p \mapsto \mathbb{E}_{\mathbb{Q}}[Y|P = p]$ is non-decreasing, then $\mathbb{E}_{\mathbb{Q}}[Y|P] = \pi^*(P)$ almost surely and $D(\mathbb{Q})$ equals

$$\max_{\pi\colon [0,1]\to[0,1]} \mathbb{E}_{\mathbb{Q}}\left[\log\left((\pi(P)/P)^Y((1-\pi(P))/(1-P))^{1-Y}\right)\right],$$

which follows by applying, in the expression above, the tower property of conditional expectations and strict concavity of $p \mapsto p\log(p) + (1-p)\log(1-p)$. Hence, in that case our test is asymptotically growth rate optimal in the sense that

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\mathbb{Q}}[\log(E_{\mathrm{HL},n,\mathrm{sym}})]}{n} \geq D(\mathbb{Q})$$

is maximal among the growth rates of all tests for the hypothesis of calibration. In the case when $p \mapsto \mathbb{E}_{\mathbb{Q}}[Y|P = p]$ is not increasing, our test is still optimal among all tests with non-decreasing alternative probabilities $\pi\colon [0,1] \to [0,1]$, by definition of the optimal isotonic approximation $\pi^*$. How large the difference in growth rate to the optimal test is depends on how strongly the isotonicity assumption is violated, and is difficult to quantify in general.

Apart from the asymptotic growth rate, the power of the test also depends on the regret, which is of order $O(n^{1/2}\log(n))$ for the algorithm presented. With a more complex exponential weights strategy given in Section 7.1 of [21], instead of the method in (2.7), one can achieve a smaller regret of order $O(n^{1/3}\log(n)^{2/3})$, which matches the optimal rate up to logarithmic factors. To see that this indeed yields a valid test, notice that the hypothesis $\mathcal{H}_{\mathrm{HL},n}$ is about the conditional expectations of $Y_i$ given $P_i$ and, hence, one can assume that $P_1, \ldots, P_n$ are given in advance to the learner. In particular, the probabilities $q_{\sigma,\sigma(i)}$ then also depend on $P_{\sigma(i+1)}, \ldots, P_{\sigma(n)}$, but they do not depend on $Y_{\sigma(i+1)}, \ldots, Y_{\sigma(n)}$. This is not allowed in the online permutation isotonic regression setting, where the learner has to make the prediction for $Y_{\sigma(i)}$ without knowledge of $P_{\sigma(i+1)}, \ldots, P_{\sigma(n)}$.

Part (iii) of Theorem 2.1 only gives a diverging lower bound on the expected value of $\log(E_{\mathrm{HL},n,\mathrm{sym}})$. However, under assumption (2.8) the average growth rate

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\pi^*(P_i)}{P_i}\right)^{Y_i}\left(\frac{1-\pi^*(P_i)}{1-P_i}\right)^{1-Y_i}$$

satisfies the strong law of large numbers, and since $D(\mathbb{Q}) > 0$ for $\mathbb{Q} \notin \mathcal{H}_{\mathrm{HL},1}$, this implies that $E_{\mathrm{HL},n,\mathrm{sym}} \to \infty$ almost surely as $n \to \infty$. In particular, for any Type-I error $\alpha$ and desired power $1-\beta$, there exists a sample size $N$ such that $\mathbb{Q}(E_{\mathrm{HL},N,\mathrm{sym}} \geq 1/\alpha) \geq 1-\beta$ for $N \geq n$.

*Proof of Theorem 2.1.* Part (i) is a consequence of the following facts. If $(\pi_n)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{F}_{\uparrow,[0,1]}$ such that $\lim_{n\to\infty} R_{\mathbb{Q}}(\pi_n) = \sup_{\pi\in\mathcal{F}_{\uparrow,[0,1]}} R(\pi)$, then by Helly's selection theorem, there exists a subsequence $(\pi_{n_k})_{k\in\mathbb{N}}$ converging pointwise to some $\pi^* \in \mathcal{F}_{\uparrow,[0,1]}$. The function $\pi(P) \mapsto \log\left((\pi(P)/P)^Y((1-\pi(P))/(1-P))^{1-Y}\right)$ inside the expectation in the definition of $R_{\mathbb{Q}}$ is strictly concave, and the set $\mathcal{F}_{\uparrow,[0,1]}$ is convex. Hence $R_{\mathbb{Q}}(\pi^*) \geq R_{\mathbb{Q}}(\pi)$ for all $\pi \in \mathcal{F}_{\uparrow,[0,1]}$, and equality holds if and only if $\pi = \pi^*$ $\mathbb{Q}$-almost-surely, provided that $R_{\mathbb{Q}}(\pi^*)$ is finite, which is shown below.

The nonnegativity in part (ii) holds because $\mathcal{F}_{\uparrow,[0,1]}$ contains the identity function, and we only have to prove that $D(\mathbb{Q}) > 0$ if $\mathbb{Q} \notin \mathcal{H}_{\mathrm{HL},1}$. For this, it is sufficient to show that there exists some $\pi$ with $R_{\mathbb{Q}}(\pi) > 0$. We start with some results about the existence of certain expected values. Since $Y|P$ is Bernoulli with expectation $\bar{\pi}(P)$, we have

$$\begin{aligned}&\mathbb{E}_{\mathbb{Q}}[\log(\pi^*(P)/P)^Y((1-\pi^*(P))/(1-P))^{1-Y} \mid P]\\ &= \bar{\pi}(P)\log(\pi^*(P)/P) + (1-\bar{\pi}(P))\log((1-\pi^*(P))/(1-P))\end{aligned}$$

and the nonnegativity of the Kullback-Leibler divergence implies

$$\begin{aligned}&\bar{\pi}(P)\log(\pi^*(P)/P) + (1-\bar{\pi}(P))\log((1-\pi^*(P))/(1-P))\\ &\leq \bar{\pi}(P)\log(\bar{\pi}(P)/P) + (1-\bar{\pi}(P))\log((1-\bar{\pi}(P))/(1-P)),\end{aligned}$$

hence we obtain

$$\begin{aligned}0 &\leq \mathbb{E}_{\mathbb{Q}}[\log(\pi^*(P)/P)^Y((1-\pi^*(P))/(1-P))^{1-Y}]\\ &\leq \mathbb{E}_{\mathbb{Q}}[\bar{\pi}(P)\log(\bar{\pi}(P)/P)\\ &\qquad + (1-\bar{\pi}(P))\log((1-\bar{\pi}(P))/(1-P))]\\ &\leq \mathbb{E}_{\mathbb{Q}}[|\log(P)| + |\log(1-P)|]\\ &\leq \sqrt{\mathbb{E}_{\mathbb{Q}}[\log(P)^2]} + \sqrt{\mathbb{E}_{\mathbb{Q}}[\log(1-P)^2]} < \infty. \qquad (2.9)\end{aligned}$$

Let now $\tilde{\pi}(P)$ be a version of the conditional expectation of $\bar{\pi}(P)$ with respect to the sigma lattice generated by $P$, which $\tilde{\pi}(P)$ satisfies the following properties:

$$\tilde{\pi} \text{ is increasing;} \qquad (2.10)$$

$$\mathbb{E}_{\mathbb{Q}}[(\bar{\pi}(P) - \tilde{\pi}(P))h(P)] \leq 0 \text{ for all increasing } h \text{ such that } \mathbb{E}_{\mathbb{Q}}[h(P)^2] < \infty; \qquad (2.11)$$

$$\mathbb{E}_{\mathbb{Q}}[\bar{\pi}(P)\mathbb{1}_B(P)] = \mathbb{E}_{\mathbb{Q}}[\tilde{\pi}(P)\mathbb{1}_B(P)] \text{ for all } B \text{ in the } \sigma\text{-field generated by } \tilde{\pi}. \qquad (2.12)$$

Equation (2.10) holds by definition of the conditional expectation given a sigma lattice, and (2.11) and (2.12) are by [5, Equations (3.9) and (3.11)]. By (2.12), we have $\bar{\pi}(P) = P$

almost surely if and only if $\tilde{\pi}(P) = P$ almost surely. By definition of $\pi^*(P)$, we know that

$$\begin{aligned}&\mathbb{E}_{\mathbb{Q}}[\log\left(\tilde{\pi}(P)/P\right)^Y((1-\tilde{\pi}(P))/(1-P))^{1-Y}]\\&\quad\leq \mathbb{E}_{\mathbb{Q}}[\log\left(\pi^*(P)/P\right)^Y((1-\pi^*(P))/(1-P))^{1-Y}].\end{aligned}$$

The goal is now to prove

$$\mathbb{E}_{\mathbb{Q}}\left[\log\left(\left(\tilde{\pi}(P)/P\right)^Y((1-\tilde{\pi}(P))/(1-P))^{1-Y}\right)\right]\geq 0,$$

with equality if and only if $\tilde{\pi}(P) = P$ $\mathbb{Q}$-almost-surely. Notice that $\tilde{\pi}$ is only defined on the support of $P$, but one can assume without loss of generality that it is defined on the whole interval $[0,1]$ by right-continuous constant extrapolation in parts where it is not defined. Since $\tilde{\pi}$ is increasing, there exist at most countably many disjoint intervals $A_i \subseteq [0,1]$, $i \in \mathcal{I}$, on which $\tilde{\pi}$ is constant with some value $c_i \in [0,1]$. Furthermore, there are at most countably many disjoint intervals $B_j$, $j \in \mathcal{J}$, whose union equals $[0,1] \setminus \bigcup_{i\in\mathcal{I}} A_i$. We now make a few case distinctions.

Fix $i$ with $c_i > 0$ and assume that $A_i = [a_i, b_i]$; the following arguments can be easily modified for the case that $A_i$ is (half-)open. Define the function

$$h_i(x) = \begin{cases}\log(1/a_i)+1, & \text{if } x < a_i,\\ \log(1/x), & \text{if } x\in[a_i,b_i],\\ -1, & \text{if } x > b_i.\end{cases}$$

Then $h_i(P)$ is square integrable due to (2.9), $h_i$ is decreasing, and constant outside of $[a_i, b_i]$, so that

$$\begin{aligned}0 &\overset{(2.11)}{\leq} \mathbb{E}_{\mathbb{Q}}[(\bar{\pi}(P)-\tilde{\pi}(P))h_i(P)]\\&\overset{(2.12)}{=} \mathbb{E}_{\mathbb{Q}}[(\bar{\pi}(P)-\tilde{\pi}(P))\log(1/P)\mathbb{1}_{A_i}(P)]\\&\overset{(2.12)}{=} \mathbb{E}_{\mathbb{Q}}[(\bar{\pi}(P)-\tilde{\pi}(P))\log(c_i/P)\mathbb{1}_{A_i}(P)]\\&= \mathbb{E}_{\mathbb{Q}}[(\bar{\pi}(P)-\tilde{\pi}(P))\log(\tilde{\pi}(P)/P)\mathbb{1}_{A_i}(P)],\end{aligned}$$

where the last step is due to the fact that $\tilde{\pi}(P) = c_i$ for $P \in A_i$. Hence

$$\begin{aligned}&\mathbb{E}_{\mathbb{Q}}[\bar{\pi}(P)\log(\tilde{\pi}(P)/P)\mathbb{1}_{A_i}(P)]\\&\quad\geq \mathbb{E}_{\mathbb{Q}}[\tilde{\pi}(P)\log(\tilde{\pi}(P)/P)\mathbb{1}_{A_i}(P)].\end{aligned}$$

If $c_i = 0$, the above inequality is still true because (2.12) implies that $\tilde{\pi}(P) = \bar{\pi}(P) = 0$ for $P \in A_i$ in that case, and we define $0\log(0) := 0$.

Similarly, for $c_i < 1$ we define

$$h_i(x) = \begin{cases}\log(1/(1-b_i))+1, & \text{if } x > b_i,\\ \log(1/(1-x)), & \text{if } x\in[a_i,b_i],\\ -1, & \text{if } x < a_i,\end{cases}$$

which is square integrable and increasing. As before,

$$\begin{aligned}0 &\overset{(2.11)}{\leq} \mathbb{E}_{\mathbb{Q}}[(\tilde{\pi}(P)-\bar{\pi}(P))h_i(P)]\\&\overset{(2.12)}{=} \mathbb{E}_{\mathbb{Q}}[(\tilde{\pi}(P)-\bar{\pi}(P))\log(1/(1-P))\mathbb{1}_{A_i}(P)]\\&\overset{(2.12)}{=} \mathbb{E}_{\mathbb{Q}}[(\tilde{\pi}(P)-\bar{\pi}(P))\log((1-c_i)/(1-P))\mathbb{1}_{A_i}(P)]\\&= \mathbb{E}_{\mathbb{Q}}[(1-\bar{\pi}(P)-(1-\tilde{\pi}(P)))\\&\qquad\times\log((1-\tilde{\pi}(P))/(1-P))\mathbb{1}_{A_i}(P)],\end{aligned}$$

so we obtain

$$\begin{aligned}&\mathbb{E}_{\mathbb{Q}}[(1-\bar{\pi}(P))\log((1-\tilde{\pi}(P))/(1-P))\mathbb{1}_{A_i}(P)]\\&\quad\geq \mathbb{E}_{\mathbb{Q}}[(1-\tilde{\pi}(P))\log((1-\tilde{\pi}(P))/(1-P))\mathbb{1}_{A_i}(P)],\end{aligned}$$

which also holds if $\tilde{\pi}(P) = 1$ on $A_i$. Hence we have shown that

$$\begin{aligned}0 &\leq \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{A_i}(P)(\tilde{\pi}(P)\log(\tilde{\pi}(P)/P)\\&\qquad+(1-\tilde{\pi}(P))\log((1-\tilde{\pi}(P))/(1-P)))]\\&\leq \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{A_i}(P)(\bar{\pi}(P)\log(\tilde{\pi}(P)/P)\\&\qquad+(1-\bar{\pi}(P))\log((1-\tilde{\pi}(P))/(1-P)))]\\&= \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{A_i}(P)\log\left(\tilde{\pi}(P)/P\right)^Y((1-\tilde{\pi}(P))/(1-P))^{1-Y}],\end{aligned}$$

and equality holds if and only if $\tilde{\pi}(P) = P$ $\mathbb{Q}$-almost-surely on $A_i$, since the Kullback-Leibler divergence is non-negative.

Consider now an interval $B_j$. Since $\tilde{\pi}$ is strictly increasing on $B_j$, the sigma field generated by $\tilde{\pi}$ contains all Borel sets which are subsets of $B_j$. Then (2.12) implies that $\tilde{\pi}(P) = \bar{\pi}(P)$ $\mathbb{Q}$-almost-surely on $B_j$, hence

$$\begin{aligned}&\mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{B_j}(P)(\bar{\pi}(P)\log(\tilde{\pi}(P)/P)\\&\qquad+(1-\bar{\pi}(P))\log((1-\tilde{\pi}(P))/(1-P)))]\\&= \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{B_j}(P)(\tilde{\pi}(P)\log(\tilde{\pi}(P)/P)\\&\qquad+(1-\tilde{\pi}(P))\log((1-\tilde{\pi}(P))/(1-P)))]\geq 0\end{aligned}$$

with equality if and only if $\tilde{\pi}(P) = P$ $\mathbb{Q}$-almost-surely on $B_j$.

With the above derivations, we obtain that for any finite number of indices $i_1,\ldots,i_n \in \mathcal{I}$, $j_1,\ldots,j_n \in \mathcal{J}$ and

$$C_n = \left(\bigcup_{k=1}^n A_{i_k}\right)\cup\left(\bigcup_{l=1}^n B_{j_l}\right),$$

the following inequalities hold,

$$\begin{aligned}0 &\leq \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{C_n}(P)(\tilde{\pi}(P)\log(\tilde{\pi}(P)/P)\\&\qquad+(1-\tilde{\pi}(P))\log((1-\tilde{\pi}(P))/(1-P)))] \quad (2.13)\\&\leq \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{C_n}(P)(\bar{\pi}(P)\log(\tilde{\pi}(P)/P)\\&\qquad+(1-\bar{\pi}(P))\log((1-\tilde{\pi}(P))/(1-P)))]. \quad (2.14)\end{aligned}$$

Since the integrand in (2.13) is non-negative and the integrand in (2.14) dominated pointwise by

$$M(P) = \bar{\pi}(P)\log(\bar{\pi}(P)/P) + (1-\bar{\pi}(P))\log((1-\bar{\pi}(P))/(1-P))$$

with $\mathbb{E}_{\mathbb{Q}}[M(P)] < \infty$, we can choose index sequences such that $\bigcup_{n=1}^N C_n$ increases to $[0,1]$, and apply Fatou's Lemma and the dominated convergence theorem to obtain

$$0 \leq \mathbb{E}_{\mathbb{Q}}[\tilde{\pi}(P)\log(\tilde{\pi}(P)/P)$$

$$+ (1-\tilde{\pi}(P))\log((1-\tilde{\pi}(P))/(1-P))] \quad (2.15)$$
$$\leq \mathbb{E}_{\mathbb{Q}}[\bar{\pi}(P)\log(\tilde{\pi}(P)/P) + (1-\bar{\pi}(P))\log((1-\tilde{\pi}(P))/(1-P))].$$

Equality in (2.15) holds if and only if $\tilde{\pi}(P) = P$ almost surely.

For part (iii), the inequality of arithmetic and geometric mean implies that

$$E \geq \left( \prod_{\sigma\in\mathcal{S}_n} \prod_{i=1}^{n} \frac{q_{\sigma,\sigma(i)}^{Y_{\sigma(i)}}(1-q_{\sigma,\sigma(i)})^{1-Y_{\sigma(i)}}}{P_{\sigma(i)}^{Y_{\sigma(i)}}(1-P_{\sigma(i)})^{1-Y_{\sigma(i)}}} \right)^{1/n!}$$
$$= \exp\left( \frac{1}{n!} \sum_{\sigma\in\mathcal{S}_n} \sum_{i=1}^{n} \log \frac{q_{\sigma,\sigma(i)}^{Y_{\sigma(i)}}(1-q_{\sigma,\sigma(i)})^{1-Y_{\sigma(i)}}}{P_{\sigma(i)}^{Y_{\sigma(i)}}(1-P_{\sigma(i)})^{1-Y_{\sigma(i)}}} \right).$$

The term inside the exponential can be written as

$$L = \mathbb{E}_\sigma \left[ \sum_{i=1}^{n} \log \frac{q_{\sigma,\sigma(i)}^{Y_{\sigma(i)}}(1-q_{\sigma,\sigma(i)})^{1-Y_{\sigma(i)}}}{P_{\sigma(i)}^{Y_{\sigma(i)}}(1-P_{\sigma(i)})^{1-Y_{\sigma(i)}}} \right],$$

which is the negative of the entropic loss defined in Section 4.4 of [20], and the expectation $\mathbb{E}_\sigma[\cdot]$ is with respect to the uniform distribution over all permutations $\sigma$ of $\{1,\dots,n\}$. It follows from Lemma 2.1, Theorem 4.3 and the proof of Theorem 4.1 of [20] that for all $K \in \mathbb{N}$,

$$L - \sum_{i=1}^{n} \log \frac{\hat{\pi}_i^{Y_i}(1-\hat{\pi}_i)^{1-Y_i}}{P_i^{Y_i}(1-P_{\sigma(i)})^{1-Y_i}} \geq -\sum_{k=1}^{n} \left( \frac{2}{K} + \frac{4K}{k}\log(1+k) \right),$$

where $\hat{\pi}_1,\dots,\hat{\pi}_n$ is the isotonic regression of $Y_1,\dots,Y_n$ on $P_1,\dots,P_n$, i.e. the maximizer of

$$(g_1,\dots,g_n) \mapsto \hat{R}(g_1,\dots,g_n;\, P_1,\dots,P_n,Y_1,\dots,Y_n),$$

as defined at (2.6). The result now follows because

$$\hat{R}(\hat{\pi}_1,\dots,\hat{\pi}_n;\, P_1,\dots,P_n,Y_1,\dots,Y_n) \geq \hat{R}(\pi^*(P_1),\dots,\pi^*(g_n);\, P_1,\dots,P_n,Y_1,\dots,Y_n)$$

and $\sum_{k=1}^{n}(2/K + 4K\log(1+k)/k) = \mathcal{O}(\sqrt{n(\log(n))^2})$ for $K$ of order $\sqrt{n/(\log(n))^2}$. □

*Remark.* A alternative idea for constructing an e-value for $\mathcal{H}_{\mathrm{HL},n}$ could be a Bayesian approach inspired by the original procedure of the HL-test. Conditional on $P_1,\dots,P_n$, the likelihood of $Y_1,\dots,Y_n$ is fully specified. For the likelihood under the alternative, one could put a meta-prior on the number $g$ of quantile based bins. Conditional on $g$, the bin probabilities are then given by a Dirichlet prior. If the hyper-parameters of the Dirichlet prior are chosen independently of the data, then the resulting likelihood ratio (e-value) does not depend on the ordering of the data. However, if one desires to choose the hyper-parameters in a data-adaptive manner, then a similar procedure with averaging over permutations as in our proposed e-value, or a repeated sample splitting approach seem necessary to avoid instabilities due to data ordering in the iid case. We do not expect to obtain universal power guarantees with this approach since it is based on binning and at least some continuity assumptions on the distribution under the alternative seem necessary.

## 2.5 A Feasible Version of the Test

The ideal test described in Algorithm 1 cannot be implemented for practically relevant $n$, as it requires to compute e-values over all $n!$ permutations of $\{1,\dots,n\}$. Even for a single permutation $\sigma$, the inner loop in Algorithm 1 has computational complexity of $\mathcal{O}(n^2)$: it requires computing $2n$ isotonic regressions to generate out-of-sample predictions. We suggest to address these problems above by the simplified version in Algorithm 2, which can be regarded as a version of the split likelihood ratio test by [40].

**Algorithm 2** Split LRT version of the e-value.

1: **Parameters:** split fraction $s \in (0,1)$, number of splits $B \in \mathbb{N}$.
2: $E_{\mathrm{HL},n} \leftarrow 0$
3: **for** $b = 1,\dots,B$ **do**
4: randomly select $\lfloor ns \rfloor$ pairs $(Y_i,P_i)$, $j \in S_b = \{i_1,\dots,i_{\lfloor ns \rfloor}\}$, without replacement
5: estimate the isotonic of regression of $(Y_i,P_i)$, $i \in S_b$, by maximizing (2.6)
6: generate predictions $q_i$ for $\mathbb{E}[Y|P_i]$, $i \in \{1,\dots,n\} \setminus S_b$, from the isotonic regression
7: $E_{\mathrm{HL},n} \leftarrow E_{\mathrm{HL},n} + \prod_{i\in\{1,\dots,n\}\setminus S_b}(q_i/P_i)^{Y_i}((1-q_i)/(1-P_i))^{1-Y_i}/B$
8: **end for**
9: **return** $E_{\mathrm{HL},n}$

A delicate point in Algorithm 2 is Step 6, where one needs to generate out-of-sample predictions from the isotonic regression fit. Naive extrapolation approaches could lead to predicted probabilities $q_i \in \{0,1\}$ and hence an e-value of zero if either $q_i = 0$ and $Y_i = 1$ or $q_i = 1$ and $Y_i = 0$.

Let $p_1 < \cdots < p_m$ denote the distinct values of $P_i$, $i \in S_b$, and $\hat{\pi}_1 \leq \cdots \leq \hat{\pi}_m$ the corresponding values of the isotonic regression. A well known result about isotonic regression states that there exists a partition of $S_b$ into index sets $\mathcal{I}_1,\dots,\mathcal{I}_d$ such that $\hat{\pi}_j$ is the empirical mean of the $Y_i$ with indices in $\mathcal{I}_j$,

$$\hat{\pi}_j = \frac{1}{\#\mathcal{I}_j} \sum_{i\in\mathcal{I}_j} Y_i.$$

To remedy the problem of predictions in $\{0,1\}$, we propose to apply the smoothed Laplace predictor, equivalent to Jef-

freys' prior in binomial proportion estimation,

$$\check{\pi}_j = \frac{1}{\#\mathcal{I}_j + 1}\left(0.5 + \sum_{i\in\mathcal{I}_j} Y_i\right) \in (0,1).$$

For out-of-sample predictions at $P_i \notin \{p_1, \dots, p_m\}$, one can then apply linear interpolation

$$q_i = \begin{cases} \dfrac{p_l - P_i}{p_l - p_k}\check{\pi}_k + \dfrac{P_i - p_k}{p_l - p_k}\check{\pi}_l, & \text{if } P_i \in [p_k, p_l], \\ \check{\pi}_1, & \text{if } P_i < p_1, \\ \check{\pi}_m, & \text{if } P_i > p_m, \end{cases}$$

where it is now guaranteed that $q_i \in (0,1)$.

## 3. SIMULATIONS

This section evaluates the empirical performance of the feasible version of the proposed test in Section 2.5 together with sensible values of the splitting fraction $s \in (0,1)$. We follow the simulation setup of [17] with a quadratic misspecification in assessing HL-type tests, which is, if at all, just slightly modified in more recent contributions [16, 42, 1, 6, 25]. Replication material for the simulations and the application in Section 4 in the statistical software R is available under https://github.com/marius-cp/eHL.

For $i = 1, \dots, 2n$ with $n \in \{1024, 2048, 4096, 8192\}$, we simulate the iid covariate $X_i \overset{\text{iid}}{\sim} \mathrm{U}(-3,3)$ and let the response variables $Y_i \sim \text{Bernoulli}(\pi_i)$ be independent, where the true conditional event probability $\bar{\pi}_i$ follows a logistic transformation of a quadratic model

$$\begin{aligned} \bar{\pi}_i = \bar{\pi}(X_i) &= \mathbb{P}(Y_i = 1 \mid X_i;\, \beta_0, \beta_1, \beta_2) \\ &= \frac{\exp(\beta_0 + \beta_1 X_i + \beta_2 X_i^2)}{1 + \exp(\beta_0 + \beta_1 X_i + \beta_2 X_i^2)}. \end{aligned} \tag{3.1}$$

We split the simulated data into an estimation set and validation set, both of size $n$. Based on the data in the estimation set, we estimate the parameters of a linear, and hence misspecified, logistic regression model by maximum likelihood and denote the parameter estimates by $(\widehat{\beta}_0, \widehat{\beta}_1)$. The probability of a positive outcome is then predicted by

$$P_i = \frac{\exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 X_i)}. \tag{3.2}$$

We vary the severity of the misspecification, expressed through the magnitude of $\beta_2$. Following [17], we characterize the "lack of linearity" through the conditions $\bar{\pi}(-3) = j - 0.00733745$, $\bar{\pi}(-1.5) = 0.05$ and $\bar{\pi}(3) = 0.95$ such that the value $j = 0$ results in the very accurate approximation $\beta_2 \approx 0$, i.e., a linear effect of $X_i$ on the log odds-ratio. We consider a sequence of 51 equally spaced values of $j$ in the interval $[0, 0.1]$. Notice that for each choice of $j$, the values of $\beta_0$ and $\beta_1$ are also determined by these conditions.

*Table 1. Rejection rates in percentage points of the classical HL test and our eHL test under the null hypothesis with $j = 0$ and the true regression parameters $(\beta_0, \beta_1)$ in (3.2) at a significance level of 5%. We treat an e-value above 20 as a rejection in the eHL test.*

| | HL | eHL | | |
|---|---|---|---|---|
| $s$ | | 1/3 | 1/2 | 2/3 |
| $n = 1024$ | 6.2 | 0.5 | 1.0 | 0.4 |
| $n = 2048$ | 5.0 | 0.1 | 0.4 | 0.6 |
| $n = 4096$ | 4.7 | 0.2 | 0.4 | 0.6 |
| $n = 8192$ | 4.5 | 0.0 | 0.1 | 0.5 |

Table 1 reports rejection rates of the tests over 1000 simulation replications, where we set $\beta_2 = 0$ (i.e., $j = 0$), and use the true regression parameters $(\beta_0, \beta_1)$ in (3.2) to guarantee that the null hypothesis $\mathcal{H}_{\text{HL},n}$ holds. For the classical HL test, we use ten equally populated (quantile-spaced) bins, where the exact procedure follows the method $\mathsf{Q}^R$ described in Appendix A. For the feasible eHL test of Section 2.5, we use the splitting fractions $s \in \{1/3, 1/2, 2/3\}$. To limit computation time, we choose a relatively low amount of bootstrap replications $B = 10$ in the eHL test as we are mainly interested in rejection rates averaged across simulation replications, and hence, stability of the test is less of a concern as e.g., in the subsequent empirical application. Here and in the following, we treat e-values above 20 as a test rejection at the 5% significance level. The table shows that all tests are well sized, where all eHL versions exhibit rejection frequencies much below the nominal value of 5%, which is not unusual for tests based on e-values.

Figure 1 analyzes the tests' behaviour under the alternative hypotheses induced by $j > 0$. Notice that we use the true parameters $(\beta_0, \beta_1)$ in (3.2) for $j = 0$ but estimates $(\widehat{\beta}_0, \widehat{\beta}_1)$ for any $j > 0$ as the pseudo-true parameters are unknown under model misspecification. In this analysis, we further include an oracle version of the eHL test, whose e-values are optimal in the sense that they are based on $q_i = \bar{\pi}_i$, i.e., the practically unknown true conditional event probabilities. The oracle eHL version with $s = 1/2$ facilitates a fair comparison with the feasible HL test based on the same splitting factor. The left panel of the figure shows classical test rejection rates for a nominal significance level of 5%. Following the explanations in Section 2.4 together with [13] and [31], a suitable measure of power for e-values is the growth rate $\mathbb{E}\big[\log(E_{\text{HL},n})\big]$, which is shown in the right panel of Figure 1, where we approximate the expectation by the average log e-value over the simulation replications. We restrict attention to $n \in \{1024, 4096\}$ as the other sample sizes do not yield further insights.

We find that all tests develop power for increasing misspecification $j$. E.g., the feasible eHL versions already have substantial power for both sample sizes against an alternative with $j \approx 0.043$, which [17] interpret as a value inducing

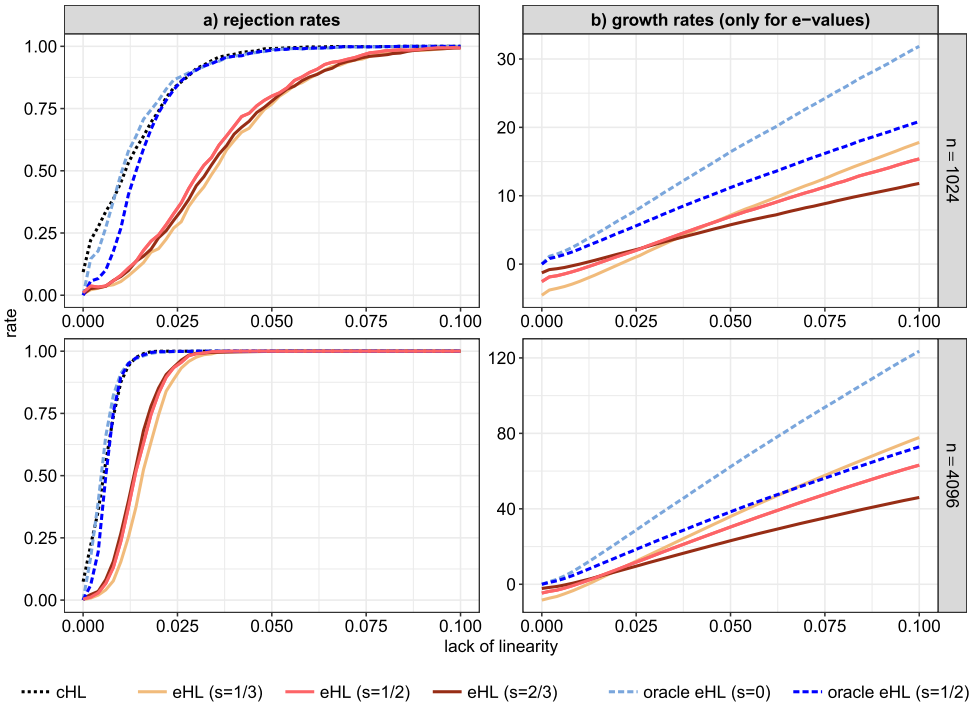


Figure 1: Rejection (left) and e-value growth (right) rates for the classical HL (cHL) test, the feasible eHL test and an oracle eHL test for a range of splitting factors $s$. The oracle eHL test is based on the true $\pi_i$. The $x$-axis contains the severity of model misspecification, and the vertically aligned plots correspond to different sample sizes.

only 'slight' misspecification. (Notice that $j \approx 0.043$ equals 0.05 in their parametrization.) There seems to be little difference among the feasible eHL tests when using different splitting fractions $s$, and hence, we do not find arguments to deviate from the natural choice of $s = 1/2$, which we continue to use in the application.

The higher power of the classical HL test can be explained by the required sample split in the eHL test, and the estimation error in assigning suitable values for $q_i$. The two oracle eHL tests make these steps redundant and hence achieve comparable power to the classical HL test. Perhaps surprisingly, the difference between the two oracle eHL tests with different $s$ is smaller than the respective difference to the feasible test versions based on estimated $q_i$'s, which means that tuning the test to a specific alternative through the $q_i$'s is the main empirical challenge of the eHL test.

Notice that the often overlooked bin specification in the classical HL test implicitly determines the set of alternatives the test has power against as e.g. illustrated in [8, Section 5]. As the sample split in the eHL test allows for *estimating* a suitable alternative, Theorem 2.1 shows that the (ideal version of the) eHL test has power against all alternatives. This power guarantee together with the eHL test's stability come at the cost of a lower power compared to the classical HL test in specific smooth forms of misspecifications as shown in Figure 1.

Turning to the growth rates of the feasible eHL tests, we find that larger choices of $s$ perform better for slight model misspecifications (small $j$) while the opposite is true for large misspecifications. This can be explained since as discussed around (2.2)–(2.3), $\bar{\pi}_i$ must be on the 'correct' side of $P_i$ to gain power, which might be violated for small $s$ (and $n$) under slight misspecifications.

## 4. APPLICATION: CREDIT CARD DEFAULTS IN TAIWAN

In this application, we analyze (re-)calibration of probability predictions for the binary event of credit card defaults in Taiwan in the year 2005. In that time period, banks in Taiwan over-issued credit cards, also to unqualified applicants, who at the same time overused the cards for consumption, resulting in severe credit card debts and damaged consumer finance confidence [43, 24]. This crisis calls for improved and in particular calibrated probability predictions for credit card defaults that can be used for a thorough risk management and improved financial regulations.

For our analysis, we use a data set of $m = 30\,000$ credit card holders from Taiwan in 2005, that is publicly available from the UCI Machine Learning Repository [10, 43] under https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. Specifically, the binary response variable $Y_i \in \{0, 1\}$ contains information on whether a default payment, $Y_i = 1$, occurred for customer $i =$

*Table 2. E-values of the eHL and the range of p-values of the classical HL test, the latter stemming from 80 reasonable binning procedures as detailed in Table 3 and Appendix A.*

| Prediction method | eHL e-values | Range of HL p-values |
|---|---|---|
| Logistic model | $7.0 \cdot 10^{28}$ | [0.00, 0.00] |
| Logistic model with increased estimation set | $9.6 \cdot 10^{22}$ | [0.00, 0.00] |
| Isotonic recalibration | 20.04 | [0.00, 0.91] |
| Bagged isotonic recalibration | 6.14 | [0.00, 0.53] |

$1, \dots, m$. We observe a relatively high rate of 22.12% of default payments in the data set that reflects the above mentioned credit card crisis. The data set further includes 23 explanatory variables, containing information on the amount of given credit, gender, education, marital status, age, and various historical payment records for the past six months.

We randomly split the data into an estimation and a Recalibration set $\mathcal{R}$ with $M = 12000$ observations each, and a Validation set $\mathcal{V}$ containing the remaining $n = 6000$ observations. We use the estimation set to fit a standard logistic regression model based on all predictor variables by maximum likelihood and compute the model predictions on the recalibration and validation sets, respectively. We run all the following tests on the validation set.

Table 2 reports the e-values of the feasible version of our calibration test described in Section 2.5 based on $B = 10000$ bootstrap replication and with a splitting factor of $s = 1/2$ that is motivated by our simulation results. We further report the range between the smallest and largest p-value of the classical HL test, where the different p-values result from five different, but natural binning procedures using $g = 5, \dots, 20$ bins, respectively. We provide further details on these implementation choices in Appendix A.

The predictions from the logistic model result in an e-value far beyond the value of 20 in Table 2, hence implying that these predictions are clearly miscalibrated. In this setting, all implementation choices of the classical HL test agree and deliver p-values very close to zero. The second row of the table shows that even when using all observations in the "increased estimation set" comprising the estimation and the recalibration set, the situation barely changes and both the eHL and HL tests agree (under all implementation choices).

As a consequence of these clear rejections, we now aim at isotonically recalibrating the probability predictions, a technique that proved valuable in other disciplines [14, 36], where it is also called "post-processing". For this, we estimate an isotonic regression on the recalibration set $\mathcal{R}$ and generate recalibrated predictions by transforming the logistic predictions on $\mathcal{V}$ with the estimated isotonic regression function. Table 2 shows that our calibration test has an e-value just above 20, i.e., a weak rejection when interpreted as a (conservative) test at the 5% level.

As a nonparametric method, the isotonic regression is known to involve substantial estimation noise that might adversely affect the recalibrated predictions. Hence, we stabilize the estimation through the classical bagging (bootstrap aggregation) method of [4]. In detail, we draw $\widetilde{B} = 100$ bootstrap samples $\mathcal{R}_b$, $b = 1, \dots, \widetilde{B}$ of size $M$ from the recalibration set $\mathcal{R}$ and estimate the isotonic regression on each bootstrap sample $\mathcal{R}_b$. The final predictions are obtained by recalibrating with the average of the estimated isotonic regression functions, displayed in Figure 2.

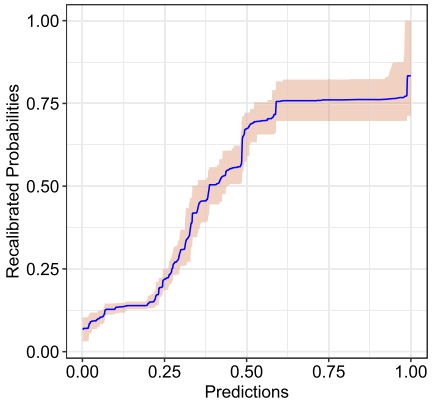


Figure 2: Bagged isotonic recalibration curve of the logit predictions. The blue curve shows the mean, and the red band the range of the pointwise 1% and 99% quantiles, over all bagging iterations.

The last row of Table 2 shows an e-value of approximately 6 implying only very weak evidence against the null hypothesis of calibration, once again illustrating the practical strength of both, bagging and recalibration methods. The estimated re-calibration function displayed in Figure 2 reinforces the importance of recalibrating the logistic model predictions by showing that it substantially deviates from the diagonal.

For these two recalibration methods, the various natural implementation choices of the HL test, further described in Appendix A, result in p-values ranging between essentially 0 and 0.91 (and 0.53 respectively). The corresponding p-value histograms in Figure 3 (and the detailed results in Table 3) show the continuum of p-values, where the null hypothesis is rejected in approximately half of the cases at the 5% level, implying that a researcher can essentially tailor the test decision to her will. As already noted by [17, 3, 22], this is

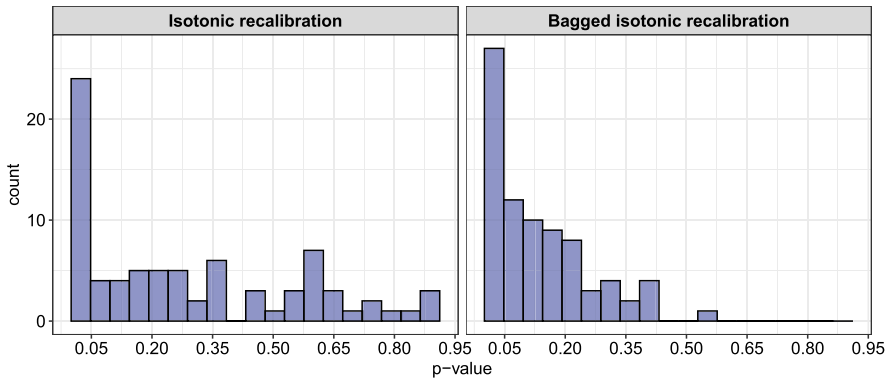


Figure 3: Histograms of $p$-values of the classical out-of-sample HL test based the five binning procedures given in Appendix A based on 5–20 bins, respectively, resulting in a total of 80 test results.

a disconcerting state of affairs for a commonly used testing procedure and calls for more robust alternatives, such as the eHL test proposed in this paper. Appendix A further shows that the feasible eHL test version is affected less by such instabilities arising from the repeated sample splits, at least if $B$ is chosen sufficiently large as in this application.

## 5. DISCUSSION

This article proposes an e-test for perfect calibration, which is a safe testing counterpart to the widely used Hosmer-Lemeshow test. The proposed eHL test follows a simple betting interpretation (see [31]) where the e-value can be seen as the factor by which we multiply the bet against the hypothesis of perfect calibration. Intuitively, when accumulating money by the bet, we gain evidence against the null. Here, the e-value depends on the probability prediction, its corresponding realization, and an arbitrary value, which we suggest estimating in a two-step approach by isotonic regression. The ideal version of the test has power against all alternatives. In order to achieve this power guarantee, it is important that for any deviation from calibration, that is, for any deviation of $P \mapsto \mathbb{E}(Y|P)$ from the identity, there is an isotonic function with strictly smaller loss. If this property can be achieved with other function classes than isotonic functions is an interesting open question.

We assess the empirical performance of the test to detect quadratic model misspecifications. The simulations show that in samples of more than 2000 observations, the eHL test allows to reliably detect levels of quadratic misspecification, which [17, p. 973] denote to be slight. The intrinsic flexibility of the e-values allows the application of stable data-driven methods (here isotonic regression) instead of the typical binning and counting technique in the HL test. However, this flexibility comes at the cost of lower power in small samples of less than 2000 observations.

The feasible version of our test is based on random splits of the training data. Since the null hypothesis $\mathcal{H}_{\mathrm{HL},n}$ requires calibration conditional on the predictions $P_1, \ldots, P_n$, one also obtains a valid test if the splits are performed systematically based on $P_1, \ldots, P_n$, for example, by choosing two subsets with similar distribution of the $P_i$ in order for the isotonic regression estimator to extrapolate well. Systematic sampling approaches to increase the power have already been applied by [11] for testing treatment effects.

Our article focuses on the batch setting where a fixed sample of size $n$ is available, rather than the online setting in which $(P_i, Y_i)$, $i \in \mathbb{N}$, arrive sequentially. However, the fact that powerful tests based on isotonic regression can be constructed in the batch setting suggests that similar approaches may be fruitful for online testing. Kotłowski et al. [21, Section 7.1] describe algorithms with sublinear regret for online isotonic regression (without the random permutation setting). We believe that in conjunction with parts (i) and (ii) of our Theorem 2.1, it is possible to derive power guarantees for sequential calibration tests where $\mathbb{E}[Y|P]$ is estimated sequentially with isotonic regression. We leave such extensions for future work.

## APPENDIX A. (IN-)STABILITY RESULTS FOR THE HL AND EHL TESTS

The classical HL test given in (1.2) is based on a partition of the unit interval into $g \in \mathbb{N}$ bins. We use the subsequently described five partitioning methods in the application in Section 4, starting with the equidistant variant:

- E: We partition the interval $\big[\min(p_i; i = 1, \ldots, n), \max(p_i; i = 1, \ldots, n)\big]$ into $g$ equidistant bins that are, apart from the first bin, open at left and closed at right.

We further use four natural implementations of "quantile-based" binning, all using a nominal number of $g$ bins. These methods mainly differ for multiple occurrences of the same forecast value, which is however not unusual in practice and is e.g., an inherent feature of methods based on decision trees or isotonic regressions.

*Table 3. p-values of the HL test based on various binning choices described in the text for the two recalibrated prediction methods from Section 4.*

| | Isotonic recalibration | | | | | Bagged isotonic recalibration | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Bins | $\mathsf{Q}^L$ | $\mathsf{Q}^R$ | $\mathsf{Q}^+$ | $\mathsf{Q}^-$ | $\mathsf{E}$ | $\mathsf{Q}^L$ | $\mathsf{Q}^R$ | $\mathsf{Q}^+$ | $\mathsf{Q}^-$ | $\mathsf{E}$ |
| 5 | 0.34 | 0.46 | 0.09 | 0.00 | 0.24 | 0.08 | 0.05 | 0.09 | 0.00 | 0.11 |
| 6 | 0.16 | 0.60 | 0.22 | 0.00 | 0.31 | 0.10 | 0.21 | 0.18 | 0.26 | 0.33 |
| 7 | 0.24 | 0.56 | 0.01 | 0.00 | 0.00 | 0.02 | 0.16 | 0.01 | 0.00 | 0.37 |
| 8 | 0.59 | 0.55 | 0.17 | 0.00 | 0.38 | 0.11 | 0.20 | 0.17 | 0.02 | 0.15 |
| 9 | 0.20 | 0.53 | 0.06 | 0.00 | 0.02 | 0.08 | 0.20 | 0.07 | 0.00 | 0.26 |
| 10 | 0.26 | 0.67 | 0.19 | 0.00 | 0.36 | 0.20 | 0.11 | 0.22 | 0.00 | 0.10 |
| 11 | 0.27 | 0.33 | 0.08 | 0.00 | 0.77 | 0.06 | 0.09 | 0.10 | 0.00 | 0.08 |
| 12 | 0.15 | 0.91 | 0.19 | 0.00 | 0.02 | 0.10 | 0.18 | 0.21 | 0.01 | 0.11 |
| 13 | 0.57 | 0.58 | 0.27 | 0.00 | 0.60 | 0.16 | 0.17 | 0.31 | 0.00 | 0.00 |
| 14 | 0.22 | 0.87 | 0.01 | 0.00 | 0.09 | 0.03 | 0.07 | 0.01 | 0.00 | 0.03 |
| 15 | 0.60 | 0.68 | 0.04 | 0.00 | 0.64 | 0.04 | 0.09 | 0.06 | 0.00 | 0.00 |
| 16 | 0.80 | 0.28 | 0.17 | 0.00 | 0.11 | 0.29 | 0.37 | 0.20 | 0.00 | 0.01 |
| 17 | 0.86 | 0.45 | 0.11 | 0.00 | 0.02 | 0.25 | 0.07 | 0.14 | 0.00 | 0.00 |
| 18 | 0.36 | 0.63 | 0.14 | 0.00 | 0.10 | 0.19 | 0.19 | 0.17 | 0.00 | 0.00 |
| 19 | 0.48 | 0.73 | 0.38 | 0.00 | 0.01 | 0.40 | 0.53 | 0.42 | 0.00 | 0.10 |
| 20 | 0.59 | 0.83 | 0.35 | 0.00 | 0.61 | 0.42 | 0.30 | 0.39 | 0.00 | 0.02 |

- $\mathsf{Q}^L$: We partition the interval $[0,1]$ into $g$ left-open and right-closed bins according to the sample quantiles (using the default `quantile()` function in `R`) at levels $1/g, \dots, (g-1)/g$. This method is denoted with the superscript $L$ as forecasts on the bin boundary are assigned to the Left bin. The first bin is also closed at left and if the sample quantiles at different levels coincide, they are ignored, resulting in possibly less than $g$ bins.
- $\mathsf{Q}^R$: As $\mathsf{Q}^L$, but we use $g$ right-open and left-closed bins such as forecasts on the bin boundaries are assigned to the Right bin.
- $\mathsf{Q}^+$: We sort the forecast-realization pairs $(p_i, y_i)_{i=1}^n$ by their forecast values $p_i$ and in the case of tied forecast values, by their realizations in *ascending* order. Based on this order, we place the observations in $g$ equally populated bins. If the size of the data set is not a multiple of $g$, excess values are redistributed in such a way that the bins with an additional observation are as far apart from each other as possible.
- $\mathsf{Q}^-$: As variant $\mathsf{Q}^+$, except that we sort in *descending* order of $y_i$ for tied forecast values.

A comparison of the methods $\mathsf{Q}^L$ and $\mathsf{Q}^R$ illustrates that assigning predictions on the bin boundaries either to the left or right bins can have consequential implications. The methods $\mathsf{Q}^+$ and $\mathsf{Q}^-$ circumvent this issue by selecting approximately equal amounts of observations into each bin, but in turn are sensitive to a change in the simple ordering of the (iid) observations in the underlying data, something that is usually ignored in applications.

While the existing literature often simply refers to "quantile-based" binning, this list shows that the HL test is sensitive to subtleties that one might easily disregard, but turn out to be consequential for the test result in some instances. This is illustrated by Table 3, which reports $p$-values for the classical HL test based on the five binning methods discussed above using $g = 5, \dots, 20$ bins respectively for the two recalibration methods used in the application in Section 4. We find that the p-values vary substantially in both, using different numbers of bins and different binning implementations. Maybe surprisingly, even for a fixed $g$, the subtleties in the four quantile-based binning choices lead to widely varying p-values.

In contrast to the classical HL test, the theoretical version of the eHL test described in Algorithm 1 is tuning parameter free due to the use of the isotonic regression method. This is unfortunately not true for the feasible eHL test described in Algorithm 2 that might be sensitive to the chosen sampling splits in the bootstrap-like replications. In particular, one has to choose the number of replications $B$ large enough such that the resulting e-values are not sensitive to the random numbers (i.e., the 'random seed') that determine the sample split.

To analyse this effect in our practical data example, Figure 4 visualizes the empirical distribution of the e-values (for tests based on different random splits), for varying bootstrap replications $B \in \{100, 500, 1000, 10000\}$. While there is indeed some variation in the test result for smaller values of $B$, the e-values are relatively stable for $B = 10000$, the choice we employ in the empirical application. E.g., for the isotonic recalibration method, essentially all e-values are between 16 and 27, implying (conservative) p-values between $1/27 \approx 0.037$ and $1/16 = 0.0625$. Similarly, the p-values in the bagged isotonic recalibration implied by the respective e-values range between $1/8 = 0.125$ and $1/4 = 0.25$. In contrast, the variation of the HL test p-values in Table 3 is

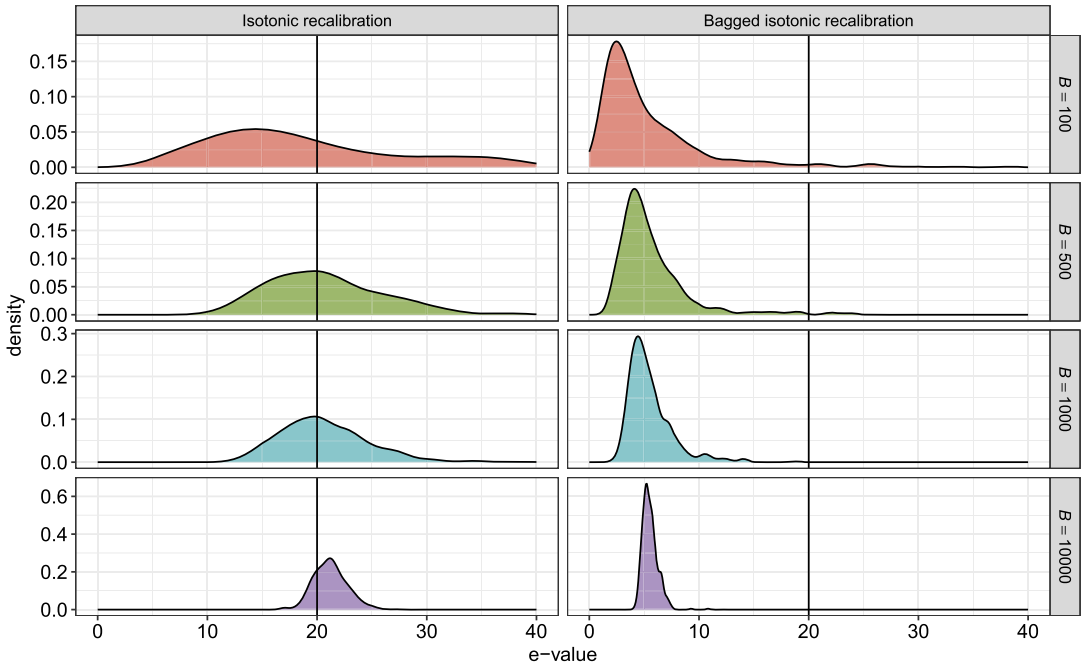


Figure 4: Kernel density estimates of 500 e-values, obtained by starting the feasible test version from different random seeds and hence implying different random splits, for the two recalibrated prediction methods in the application based on $B = \{100, 500, 1000, 10000\}$ bootstrap replications (in Algorithm 2). For the bagged recalibration model, we observe 17 e-values above 20 for $B = 100$, 5 for $B = 500$, and none for $B \in \{1000, 10000\}$.

much more substantial and includes clear test rejections as well as many p-values above any commonly chosen significance level.

## ACKNOWLEDGEMENTS

For helpful comments, we would like to thank the two guest editors as well as two anonymous referees.

## FUNDING

A. Henzi and J. Ziegel gratefully acknowledge financial support from the Swiss National Science Foundation. T. Dimitriadis gratefully acknowledges financial support from the German Research Foundation (DFG) through grant number 502572912.

*Accepted 2 December 2023*

## REFERENCES

[1] Allison, P. J. Measures of fit for logistic regression. *Paper 1485-2014, SAS Global Forum 2014, Washington DC.*

[2] Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and Silverman, E. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics* **26**(4) 641–647 (1955). https://doi.org/10.1214/aoms/1177728423. MR0073895

[3] Bertolini, G., D'Amico, R., Nardi, D., Tinazzi, A. and Apolone, G. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *Journal of Epidemiology and Biostatistics* **5**(4) 251–253 (2000).

[4] Breiman, L. Bagging predictors. *Machine Learning* **24**(5) 123–140 (1996).

[5] Brunk, H. D. Conditional expectation given a $\sigma$-lattice and applications. *Annals of Mathematical Statistics* **36**(5) 1339–1350 (1965). https://doi.org/10.1214/aoms/1177699895. MR0185629

[6] Canary, J. D., Blizzard, L., Barry, R. P., Hosmer, D. W. and Quinn, S. J. A comparison of the Hosmer–Lemeshow, Pigeon–Heyse, and Tsiatis goodness-of-fit tests for binary logistic regression under two grouping methods. *Communications in Statistics – Simulation and Computation* **46**(3) 1871–1894 (2017). https://doi.org/10.1080/03610918.2015.1017583. MR3625254

[7] Casgrain, P., Larsson, M. and Ziegel, J. Anytime-valid sequential testing for elicitable functionals via supermartingales. *Bernoulli* (2022). To appear.

[8] Dimitriadis, T., Dümbgen, L., Henzi, A., Puke, M. and Ziegel, J. Honest calibration assessment for binary outcome predictions. *Biometrika* **110**(3) 663–680 (2023). https://doi.org/10.1093/biomet/asac068. MR4627777

[9] Dimitriadis, T., Gneiting, T. and Jordan, A. I. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences* **118**(8), e2016191118 (2021). https://doi.org/10.1073/pnas.2016191118. MR4275118

[10] Dua, D. and Graff, C. UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science. Accessible at http://archive.ics.uci.edu/ml.

[11] Duan, B., Ramdas, A. and Wasserman, L. Interactive rank testing by betting. In *Conference on Causal Learning and Reasoning* **177** 201–235 (2022).

[12] Flach, P. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data.* Cambridge University Press, Cambridge, UK (2012). https://doi.org/10.1017/CBO9780511973000. MR3088204

[13] Grünwald, P., de Heide, R. and Koolen, W. Safe testing (2020). Preprint. arXiv:1906.07801. https://doi.org/10.1007/978-3-642-39091-3_21. MR3108509

[14] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q. On calibration of modern neural networks. (D. Precup and Y. W. Teh, eds.) In *Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research* **70** 1321–1330 (2017).

[15] Henzi, A. and Ziegel, J. Valid sequential inference on probability forecast performance. *Biometrika* **109**(3) 647–663 (2022). https://doi.org/10.1093/biomet/asab047. MR4472840

[16] Hosmer, D. W. and Hjort, N. L. Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine* **21**(18) 2723–2738 (2002).

[17] Hosmer, D. W., Hosmer, T., Le Cessie, S. and Lemeshow, S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* **16**(9) 965–980 (1997).

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

[19] Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X. *Applied Logistic Regression*. Wiley, Hoboken, NJ (2013). https://doi.org/10.1002/9781118548387

[20] Kotlowski, W., Koolen, W. M. and Malek, A. Random permutation online isotonic regression. In *Advances in Neural Information Processing Systems* (2017).

[21] Kotłowski, W., Koolen, W. M. and Malek, A. Online isotonic regression. In *Annual Conference on Learning Theory (COLT-16)* **49** 1165–1189 (2016).

[22] Kuss, O. Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine* **21**(24) 3789–3801 (2002).

[23] Lee, L. Y., Cazier, J.-B., Angelis, V., Arnold, R., Bisht, V., Campton, N. A., Chackathayil, J., Cheng, V. W., Curley, H. M., Fittall, M. W., Freeman-Mills, L., Gennatas, S., Goel, A., Hartley, S., Hughes, D. J., Kerr, D., Lee, A. J., Lee, R. J., McGrath, S. E., Middleton, C. P., Murugaesu, N., Newsom-Davis, T., Okines, A. F., Olsson-Brown, A. C., Palles, C., Pan, Y., Pettengell, R., Powles, T., Protheroe, E. A., Purshouse, K., Sharma-Oates, A., Sivakumar, S., Smith, A. J., Starkey, T., Turnbull, C. D., Várnai, C., Yousaf, N., Team, U. C. M. P., Kerr, R. and Middleton, G. Covid-19 mortality in patients with cancer on chemotherapy or other anticancer treatments: a prospective cohort study. *The Lancet* **395**(10241) 1919–1926 (2020).

[24] Lo, H.-Y. and Harvey, N. Shopping without pain: Compulsive buying and the effects of credit card availability in Europe and the Far East. *Journal of Economic Psychology* **32**(1) 79–92 (2011).

[25] Nattino, G., Pennell, M. L. and Lemeshow, S. Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test. *Biometrics* **76**(2) 549–560 (2020). https://doi.org/10.1111/biom.13249. MR4125279

[26] Neblett Fanfair, R., Benedict, K., Bos, J., Bennett, S. D., Lo, Y.-C., Adebanjo, T., Etienne, K., Deak, E., Derado, G., Shieh, W.-J., Drew, C., Zaki, S., Sugerman, D., Gade, L., Thompson, E. H., Sutton, D. A., Engelthaler, D. M., Schupp, J. M., Brandt, M. E., Harris, J. R., Lockhart, S. R., Turabelidze, G. and Park, B. J. Necrotizing cutaneous mucormycosis after a tornado in Joplin, Missouri, in 2011. *New England Journal of Medicine* **367**(23) 2214–2225 (2012).

[27] Orabona, F. and Jun, K.-S. Tight concentrations and confidence sequences from the regret of universal portfolio (2021). Preprint. arXiv:2110.14099.

[28] Ostrosky-Zeichner, L., Harrington, R., Azie, N., Yang, H., Li, N., Zhao, H., Koo, V. and Wu, E. Q. A risk score for fluconazole failure among patients with candidemia. *Antimicrobial Agents and Chemotherapy* **61**(5) e02091–16 (2017).

[29] Paul, P., Pennell, M. L. and Lemeshow, S. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. *Statistics in Medicine* **32**(1) 67–80 (2013). https://doi.org/10.1002/sim.5525. MR3017884

[30] Rissanen, J. and Roos, T. Conditional NML universal models. In *2007 Information Theory and Applications Workshop* 337–341 (2007).

[31] Shafer, G. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **184**. 407–431 (2021). https://doi.org/10.1111/rssa.12647. MR4255905

[32] Shafer, G. and Vovk, V. *Game-Theoretic Foundations for Probability and Finance*. Wiley, Hoboken, NJ (2019). https://doi.org/10.1002/9781118548035

[33] Shekhar, S. and Ramdas, A. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory* (2023). To appear. https://doi.org/10.1109/TIT.2023.3305867

[34] Strieder, D. and Drton, M. On the choice of the splitting ratio for the split likelihood ratio test. *Electronic Journal of Statistics* **16**(2) 6631–6650 (2022). https://doi.org/10.1214/22-ejs2099. MR4527023

[35] Tse, T. and Davison, A. C. A note on universal inference. *Stat* **11**(1), e501 (2022). https://doi.org/10.1002/sta4.501. MR4529724

[36] Vannitsem, S., Wilks, D. S. and Messner, J. *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, Amsterdam (2018).

[37] Vovk, V., Petej, I. and Fedorova, V. Large-scale probabilistic predictors with and without guarantees of validity. In *Advances in Neural Information Processing Systems* (2015).

[38] Vovk, V. and Wang, R. E-values: Calibration, combination and applications. *The Annals of Statistics* **49**(3) 1736–1754 (2021). https://doi.org/10.1214/20-aos2020. MR4298879

[39] Wang, R. and Ramdas, A. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**(3) 822–852 (2022). MR4460577

[40] Wasserman, L., Ramdas, A. and Balakrishnan, S. Universal inference. *Proceedings of the National Academy of Sciences* **117**(29) 16880–16890 (2020). https://doi.org/10.1073/pnas.1922664117. MR4242731

[41] Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2023). To appear.

[42] Xie, X.-J., Pendergast, J. and Clarke, W. Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors. *Computational Statistics & Data Analysis* **52**. 2703–2713 (2008). https://doi.org/10.1016/j.csda.2007.09.027. MR2419536

[43] Yeh, I.-C. and Lien, C.-h. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* **36**(2) 2473–2480 (2009).

[44] Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02* 694–699. Association for Computing Machinery, New York, NY, USA (2002).



Alexander Henzi. Seminar for Statistics, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland. E-mail address: alexander.henzi@stat.math.ethz.ch

Marius Puke. Institute of Economics and Computational Science Hub (CSH), University of Hohenheim, Schloss Hohenheim 1 C, 70593 Stuttgart, Germany. E-mail address: marius.puke@uni-hohenheim.de

Timo Dimitriadis. Alfred-Weber-Institute for Economics, Heidelberg University, Bergheimer Str. 58, 69115 Heidelberg, Germany. E-mail address: timo.dimitriadis@awi.uni-heidelberg.de

Johanna Ziegel. Institute of Mathematical Statistics and Actuarial Science, University of Bern, Alpeneggstrasse 22, 3012 Bern, Switzerland. E-mail address: johanna.ziegel@stat.unibe.ch