Received: 27 September 2023 | Revised: 1 February 2024 | Accepted: 25 March 2024

DOI: 10.1002/sim.10074

RESEARCH ARTICLE

Statistics in Medicine WILEY

Individualized empirical null estimation for exact tests of healthcare quality

Nicholas Hartman[1,2] | **Kevin He**[1,2]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA

[2]Kidney Epidemiology and Cost Center, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Nicholas Hartman, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA.
Email: nhar@umich.edu

Funding information
National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Number: R01DK129539

United States federal agencies evaluate healthcare providers to identify, flag, and potentially penalize those that deliver low-quality care compared to national expectations. In practice, evaluation metrics are inevitably impacted by unobserved confounding factors, which reduce flagging accuracy and cause the statistics to be overdispersed relative to the theoretical null distributions. In response to this issue, several authors have proposed individualized empirical null (IEN) methods to estimate an appropriate null distribution for each provider's evaluation statistic while taking into account the provider's effective size. However, existing IEN methods require that the statistics asymptotically follow normal distributions, which often does not hold in applications with small providers or misspecified models. In this article, we develop an IEN framework for exact hypothesis tests that accounts for the impact of unobserved confounding without making any asymptotic assumptions. Simulations show that the proposed IEN method has greater flagging accuracy compared to conventional approaches. We apply these methods to evaluate dialysis facilities and transplant centers that are monitored by the Centers for Medicare and Medicaid Services.

KEYWORDS
empirical null, health policy, healthcare provider profiling, unobserved confounding

1 | INTRODUCTION

In the United States (U.S.), healthcare providers (eg, hospitals, physicians, and transplant centers) are routinely monitored and evaluated by government regulators, such as the Centers for Medicare and Medicaid Services (CMS), to maintain standards of patient care.[1-3] A primary objective in these profiling efforts is to identify providers that are underperforming and delivering healthcare of extremely low quality. In practice, statistical tests are used to assess departures from the national expectations and flag underperforming providers so that policymakers may intervene.[4,5] Considering that flagged providers often face financial penalties and reputational damage, it is important that these statistical tests limit false-flagging rates to a modest level while maintaining high power to correctly identify inadequate care.

One major obstacle in developing hypothesis tests for provider profiling is the presence of unobserved confounding factors.[6-8] Measures of healthcare quality are usually developed from risk-adjustment models, which are fit to healthcare administrative data, and attempt to control for confounding risk factors that impact patient outcomes but are outside of

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2024 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

the providers' control. For example, even providers with high-quality care may have elevated rates of adverse outcomes if they treat many older patients with comorbidities, and providers have no control over the pre-existing characteristics of their patient populations.[9] In most applications, it is impossible to observe and adjust for all of the relevant confounding factors that impact patient outcomes, and as a result of this model misspecification, the provider-level test statistics tend to be overdispersed under the null hypotheses.[5] Thus, applications of naive hypothesis tests that ignore this overdispersion will lead to high false-flagging rates, and this issue has been shown to most severely affect providers with large effective sizes, which we define formally in Section 3.1.

In response to these challenges, several authors[8,9] have applied empirical null (EN) methods, which were originally proposed by Efron (2004) for genetics applications,[10] and use information from the data to estimate an appropriate null distribution for the test statistics and limit false-flagging rates. Most recently, flexible and "individualized" empirical null (IEN) methods have been developed that model the null variance as a function of effective provider size and allow each provider's statistic to have a unique null distribution.[8,11] In the context of healthcare provider profiling, IEN methods are needed over traditional EN methods because providers typically have very different effective sizes, and as described above, overdispersion due to unobserved confounding is known to increase with effective provider size.

To the best of our knowledge, all existing IEN methods for provider profiling rely on the assumption that the test statistics (eg, standardized Z-scores) follow asymptotic normal distributions. In practice, this assumption can be violated due to two main reasons. First, it is often necessary to evaluate providers with small effective sizes, making the use of theoretical large sample properties questionable. Second, even for very large providers, model misspecification can prevent the statistics from converging in distribution to a normal density as expected. Considering that the empirical null is intended to address the impact of model misspecification (ie, unobserved confounding) on statistical inference, among other issues, this aspect of existing IEN methods is an important limitation. Under many scenarios, the asymptotic normality assumption is inappropriate, causing conventional estimates of the IEN distributions to depart from the true null distributions, increasing false-flagging rates and decreasing true-flagging rates. Thus, new IEN methods are needed that avoid the large-sample arguments used in the current literature.[8,11]

In this article, we address the limitations of existing inference approaches by developing novel IEN methods that do not require any assumptions about the asymptotic properties of the test statistics. The proposed IEN method is designed for two of the most widely-used exact tests of provider performance, the Exact Poisson test and the Exact Poisson-Binomial test, which are based on discrete patient outcome variables (eg, counts of the number of deaths occurring among each provider's patient case-mix). Simulations show that the proposed model accurately estimates the parameters of the null distributions, and the exact IEN distributions provide superior fits to the null test statistics and better flagging properties in repeated samples, compared to naive testing methods that ignore overdispersion and existing IEN approaches that depend on large-sample arguments. By avoiding the asymptotic assumptions of conventional methods, which likely do not hold in provider profiling applications, the proposed IEN method contributes clinically-meaningful improvements in the accuracy of healthcare provider evaluations, especially when smaller providers are being evaluated or when considerable unobserved confounding is present. We apply the proposed methods to evaluate mortality outcomes in dialysis facilities and organ offer acceptance practices in kidney transplant centers.

2 | NOTATION

Let $i = 1, \ldots, I$ index the healthcare providers under evaluation, and let $j = 1, \ldots, n_i$ index the patients that are treated by the ith provider. Define a random variable Y_{ij} which represents the patient outcome being assessed (eg, hospital readmission). Following the provider profiling notation introduced in previous works,[11,12] it is often reasonable to assume that Y_{ij} arises from a generalized linear model (GLM) with canonical parameter

$$\theta^*_{ij} = \mu + \gamma^*_i + \alpha_i + \boldsymbol{X}^\top_{ij}\boldsymbol{\beta}, \tag{1}$$

where μ is the population norm, γ^*_i is the quality of care effect (ie, the treatment effect) of interest, α_i is an unobserved random confounding term that is unrelated to quality of care and has variance σ^2_α, and $\boldsymbol{X}_{ij}$ is a vector of observed patient-level risk factor variables with effects $\boldsymbol{\beta}$. In practice, α_i may represent the influence of unmeasurable variables beyond the providers' treatment, such as variation in population-level health risk factors. Furthermore, we define $g(\cdot)$ as the canonical link function (commonly $\log(\cdot)$ or $\text{logit}(\cdot)$), $b(\cdot)$ with $b'(\cdot) = g^{-1}(\cdot)$, and $a(\phi)$ as a function of the model's nuisance parameter ϕ.

The α_i term reflects the amount of risk within each provider's population that has not been accounted for through the observable $\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}$ term. Thus, the between-provider variation in these unobserved risk levels can bias the estimated association between the provider's healthcare quality (the exposure of interest) and the provider's patient health outcomes (the outcome of interest). One example of an unobserved confounding factor in the context of healthcare provider profiling is the impact of the COVID-19 pandemic on health systems, as this is difficult to measure and fully describe using administrative healthcare data.[12,13] In Supplementary Figure 1, we summarize these assumed relationships between the exposure, outcome, and confounding variables for provider profiling applications.

Because α_i is unobserved, we cannot distinguish the target parameter γ_i^* from the confounding term α_i, and the observed data only contain information to estimate the combined effect $\gamma_i^* + \alpha_i$. This misspecified model requires a constraint in order to identify the parameters, such as $\sum_{i=1}^{I}\gamma_i^* + \alpha_i = 0$. While some individual providers may have small effective sizes, the total sample sizes of disease registries for provider profiling are often extremely large. For this reason, we assume that the estimates of μ and $\boldsymbol{\beta}$ have negligible sampling variability and can be used to develop a valid inference procedure for the treatment effects $\gamma_i^*, i = 1, \ldots, I$. This assumption is widely used in the provider profiling literature, though it can be relaxed if needed for smaller samples to account for uncertainty in the effect estimates.[4-7,9]

The γ_i^* parameter in Equation (1) can be treated as either a fixed or random effect, and each of these models has important advantages and disadvantages in the context of provider profiling applications. For example, the random-effect (RE) model can suffer from biased estimates if there is correlation between the provider effects and the patient characteristics (eg, better providers treat more patients with high risk factors), which is often the case. Furthermore, the RE shrinkage estimates can be especially poor for outlying providers with extreme treatment effects, and these providers are usually of special interest to policymakers.[9] Despite these limitations, the RE model allows for additional stability and efficiency when there are providers with small sample sizes. In contrast, the fixed-effects (FE) model is frequently used for provider profiling because it achieves unbiased estimation of the treatment effects for all providers, even when the RE assumptions are violated. In this article, we present the methods under an FE framework, but one could easily define the model and test statistics under the alternative RE framework instead. It is important to note that both frameworks are susceptible to the impact of unobserved confounding from α_i in the same way. That is, as the provider size increases, the probability of false flagging due to unobserved confounding increases, so IEN methods are needed regardless of which model is chosen.[8,9]

Risk-adjusted tests of healthcare quality are often performed by comparing the number of outcomes (eg, deaths or hospital readmissions) observed within a provider's patient population to the number of outcomes expected based on national standards.[5] The expected number is a function of the national average for healthcare quality and the observed confounding risk factors within the provider's patient population (ie, the $\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}$ term in (1)). More formally, the observed number of outcomes for the ith provider is $O_i = \sum_{j=1}^{n_i} Y_{ij}$, and the expected number, which adjusts for the observable confounding variables, is $E_i = \sum_{j=1}^{n_i} b'(\theta_{ij}^0)$, where $\theta_{ij}^0 = \mu + \boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}$. For each provider, these comparisons are intended to assess the null hypothesis of $H_{0,i} : \gamma_i^* = 0$, which corresponds to a provider delivering average care that is consistent with national norms. However, as we will see in Section 3, the unobserved confounding term, α_i, introduces obstacles in conducting these tests. That is, because the $E_i, i = 1, \ldots, I$ do not account for the influence of α_i, test statistics based on simple comparisons of O_i and E_i do not fully reflect healthcare quality, and further corrections are needed to obtain accurate flagging results. For simplicity, we omit discussion of issues that are relevant to formal hypothesis testing frameworks, such as multiple testing corrections, though these methods may be considered and implemented.[7,10]

3 | EXISTING INFERENCE APPROACHES FOR PROVIDER PROFILING

3.1 | Asymptotic tests

Standardized Z-scores are commonly used to construct test statistics from the O_i and E_i values and flag low-quality providers. Under (1), a naive version of these Z-scores is typically defined as

$$Z_{FE,i} = \frac{O_i - E_i}{\sqrt{a(\phi)\tilde{n}_i}}, \tag{2}$$

where $\tilde{n}_i = \sum_{j=1}^{n_i} b''(\theta_{ij}^0)$ is the effective provider size, and the FE subscript denotes the fixed-effects score test that this statistic corresponds to, based on the misspecified null hypothesis $H_{0i} : \gamma_i^* + \alpha_i = 0$. It is important to note that this misspecified null hypothesis involves the unobserved confounding term α_i, and the $Z_{FE,i}, i = 1, \ldots, I$ values do not solely

measure the treatment effects of interest, $\gamma_i^*, i = 1, \ldots, I$.[12] Because of this, the asymptotic null distribution of $Z_{FE,i}$ is not a standard normal density as one would expect. In fact, previous works have shown that the null variance of $Z_{FE,i}$ is larger than one, and approximately a linear function of $\tilde{n}_i$[11]:

$$\text{Var}[Z_{FE,i}|\gamma_i^* = 0] \approx 1 + \varphi\tilde{n}_i, \tag{3}$$

where $\varphi = \sigma_\alpha^2 \Big/ a(\phi)$. For linear models with $g(\theta_{ij}^*) = \theta_{ij}^*$, the relationship in (3) is exact.[8]

EN estimation methods were originally developed for applications where $\tilde{n}_i$ is the same across all units. Under this setting, the null variance is estimated by assuming that all Z-scores which fall within some interval around zero correspond to "null" units with $\gamma_i^* = 0$. The marginal distribution of the naive Z-scores is then modeled by an asymptotic truncated normal density, based on the assumed "null interval," and an estimate of the null variance is obtained through maximum likelihood estimation.[10,14] In contrast, IEN methods recognize that $\tilde{n}_i$ varies widely across healthcare providers, and allow each provider's Z-score statistic to have its own null interval and variance in the likelihood function. If one naively applies traditional EN methods to situations with varying $\tilde{n}_i$, the null variance will be overestimated for some providers and underestimated for others, making the flagging results untrustworthy. Thus, IEN methods are most appropriate for provider profiling, where $\tilde{n}_i$ often has a wide range.

For existing IEN methods, the assumption that $Z_{FE,i}$ asymptotically follows a normal distribution under the null hypothesis is required in both the null variance estimation and the hypothesis test construction. The theoretical justification for using a truncated normal distribution in the likelihood function is based on the assumed asymptotic properties. Furthermore, from the estimate $\hat{\varphi}$, the modified IEN-based Z-score is typically computed as $Z_{EN,i} = Z_{FE,i}/\sqrt{1 + \hat{\varphi}\tilde{n}_i}$ and compared to a quantile of the standard normal distribution to test the null hypothesis of interest, $H_{0,i} : \gamma_i^* = 0$. This testing approach is only appropriate when the null distribution of the $Z_{FE,i}$ values has converged to an asymptotic normal density. If this is not the case, the existing IEN methods may falsely flag average providers at a high rate and also fail to detect many underperforming providers. In this article, we aim to address overdispersion due to unobserved confounding as in previous empirical null works, but we further consider the impact of small sample sizes and model misspecification on the assumed asymptotic normal density of existing IEN methods.

3.2 | Naive exact tests

Exact tests of healthcare quality are based on the assumed distribution for $Y_{ij}|\theta_{ij}^*$ in (1). Under perfect risk adjustment (ie, $\sigma_\alpha^2 = 0$), the conditional distribution of the observed outcome, given the expected outcome, $O_i|E_i$, can often be derived analytically. For example, when evaluating healthcare providers with respect to mortality, hospital readmission, or other count outcomes, it is common to model $Y_{ij}|\theta_{ij}^*$ with a Poisson distribution having mean $b'(\theta_{ij}^*) = \exp(\theta_{ij}^*)$. This implies that O_i exactly follows a Poisson(E_i) distribution under the null hypothesis.[15] The Exact Poisson test has been used in many studies of the Standardized Mortality Ratio (SMR), which is a well-known measure of healthcare provider performance.[16] However, when the risk adjustment is incomplete (ie, $\sigma_\alpha^2 > 0$), the Poisson distributional assumption is invalid, and an appropriate exact null distribution must be estimated empirically for each provider under evaluation. Guaran et al. (2018) have proposed an EN estimation strategy for the Poisson distribution,[17] but this method assumes that each unit has the same null distribution and is invalid for provider profiling applications where the effective sample sizes vary substantially, as described in Section 3.1. Thus, we do not consider this method further in this article.

3.3 | Exact *P*-value inversion tests

Previous authors have recognized that the asymptotic normal distribution is often not an appropriate null model for the naive Z-scores in Equation (2).[16] As a result, the exact tests described in Section 3.2 have become popular alternatives for healthcare quality assessments. On the other hand, since IEN methods have not been available for these exact tests, many proceed by ignoring the issue of unobserved confounding, or by applying EN methods without any individualization, and both of these approaches may severely decrease flagging accuracy (as described in Section 3.1). Others have obtained p-values from the exact tests in Section 3.2, computed pseudo-Z-scores, $Z_i = \Phi^{-1}(p\text{-value}_i)$, where Φ is the cumulative distribution function of the standard normal distribution, and applied IEN

methods to these pseudo-Z-scores based on the normality assumptions from Section 3.1.[15] This alternative approach is problematic, because under incomplete risk adjustment, the p-values do not follow a Uniform(0,1) distribution, and as a result, there is no theoretical justification for these pseudo-Z-scores to follow normal distributions. Furthermore, the formula in (3) relies on the specific form of $Z_{FE,i}$ in (2), so the assumed null variances are misspecified for the pseudo-Z-scores.

Another limitation of this method is that numerical floating point issues can arise for providers with very extreme deviations from the expected outcomes. In these cases, the p-values may be so close to zero that computers cannot store them precisely, and instead assign them to be exactly equal to zero. This behavior produces infinite Z-scores which can disrupt the IEN algorithms, and a common ad-hoc solution to this problem is to replace any p-value of zero with the smallest number allowable by the software. In Section 8.2, we show that this method may cause inaccurate results that are difficult to interpret.

4 | INDIVIDUALIZED EMPIRICAL NULL FOR EXACT POISSON TESTS

4.1 | Framework

We now develop an IEN framework for the widely-used Exact Poisson test, where $Y_{ij}|\theta^*_{ij} \sim \text{Poisson}(\theta^*_{ij})$. Under the null hypothesis of interest ($H_{0,i} : \gamma^*_i = 0$), and conditional on α_i,

$$(O_i|\alpha_i, \gamma^*_i = 0) \sim \text{Poisson}(E_i\lambda_i),$$

where $\lambda_i = \exp(\alpha_i)$. As described in Section 3, the E_i can be obtained from a risk-adjustment model that incorporates observed predictors of the outcome variable. In order to avoid the use of large-sample arguments, we must make one extra distributional assumption beyond the underlying model in (1). Specifically, we assume that $\lambda_i \sim \text{Gamma}(s, r)$. This Gamma-Poisson model is mathematically convenient, and it has been applied in other contexts of provider profiling, such as Bayesian estimation of healthcare quality measures.[6] In addition, the Gamma distribution is often used to model frailty terms in the Cox proportional hazards model, which have a similar form as λ_i in our null model.[18] Supplementary Figure 2 shows that this model is flexible enough to accommodate a wide range of shapes for the distribution of λ_i, and in Section 7 we show that the flagging rates are highly robust to this Gamma distributional assumption. Thus, we argue that the Gamma density is a suitable model for the distribution of λ_i, and our framework is not substantially more restrictive than existing EN methods for Poisson outcomes.

By integrating over α_i, we have that O_i marginally follows a Negative Binomial distribution under the null hypothesis:

$$(O_i|\gamma^*_i = 0) \sim \text{NB}\left(s, \frac{r/E_i}{1 + r/E_i}\right). \tag{4}$$

The above null distribution depends on $\tilde{n}_i = E_i$, and in particular $\text{Var}[O_i|\gamma^*_i = 0]$ is an increasing function of $\tilde{n}_i$. This result is consistent with previous works for asymptotic IEN tests (Section 3.1), which have shown that overdispersion due to unobserved confounding is more severe when $\tilde{n}_i$ is large. If the risk adjustment is perfect such that no unobserved confounding exists (ie, $\sigma^2_\alpha = 0$), then this framework reduces to the original Exact Poisson test described in Section 3.2. The derived Negative Binomial null distribution is also able to accommodate overdispersion that is due to sources other than unobserved confounding. That is, even if $\text{Var}[Y_{ij}|\theta^*_{ij}] > E[Y_{ij}|\theta^*_{ij}]$, such that the original Poisson null distribution is misspecified, the proposed Negative Binomial null distribution is flexible enough to provide a good fit to the observed outcome counts for null centers, and we show this numerically in Section 7. In Sections 4.2 and 4.3, we develop methods to estimate s and r in (4).

4.2 | Null intervals and likelihood function

In this section, we construct a likelihood function for s and r that is analogous to those used in asymptotic IEN estimation methods. Let $p_i(c; s, r)$ be the probability mass function for the $\text{NB}\left(s, \frac{r/E_i}{1+r/E_i}\right)$ distribution in (4), and define $P_i(d; s, r) =$

$\sum_{c=0}^{d} p_i(c; s, r)$. We set the null interval for the ith provider as $[A_i, B_i]$, where

$$A_i = \min\left\{d : P_i\left(d; s^{(0)}, r^{(0)}\right) > u\right\},$$

$$B_i = \min\left\{d : P_i\left(d; s^{(0)}, r^{(0)}\right) > 1 - u\right\},$$

and u is a pre-specified percentile value (such as 2.5%). The set of indices that we assume corresponds to null centers is $S_0 = \{i : O_i \in [A_i, B_i]\}$.

Using the above null intervals, and truncated Negative Binomial probability mass functions, we develop the following likelihood function for s, r, and π_0, where π_0 is the proportion of providers that delivers "average care" (ie, with $\gamma_i^* = 0$) and $Q_i(s, r) = P_i(B_i; s, r) - P_i(A_i - 1; s, r)$:

$$L(s, r, \pi_0) = \prod_{i \in S_0} \pi_0 p_i(O_i; s, r) \prod_{i \notin S_0} \{1 - \pi_0 Q_i(s, r)\}. \tag{5}$$

There is no analytic solution for the maximum likelihood estimators (MLEs) of s, r, and π_0, and we must maximize (5) using numerical methods. In Section 4.3, we propose an optimization algorithm to find the MLEs, and we derive analytic formulas for suitable initial values of s and r, which are needed to implement these numerical methods.

4.3 | Optimization and initial values

To maximize (5) and obtain MLEs for s and r, we adapt the optimization algorithm originally developed for asymptotic IEN methods.[11] For a given value of π_0, we maximize the likelihood function in (5) with respect to s and r, where π_0 is treated as a known constant. This optimization step is implemented using the Nelder-Mead algorithm.[19] Then, we apply a grid search algorithm to maximize over π_0, with the corresponding $\hat{s}$ and $\hat{r}$ plugged in to (5).

In the Nelder-Mead algorithm, initial values must be specified for s and r that are close enough to the global maximizers so that the algorithm can converge to the MLEs. For existing IEN methods that are based on asymptotic normal distributions, there is only one parameter to estimate (σ_α^2), and it is straightforward to specify an initial value for σ_α^2 using properties of the normal distribution. Specifically, robust regression methods are typically applied to estimate the marginal variance of $Z_{FE,i}$, while down-weighting the impact of outliers or "non-null" providers. This marginal variance estimate can then be used with (3) to obtain a rough estimate of σ_α^2, which serves as the initial value. For the proposed Exact IEN method, there are two parameters to estimate (s and r), and the marginal variance of the $Z_{FE,i}$ scores has no obvious connection with these parameters. To overcome this challenge, we derive a set of equations that allows us to solve for reasonable initial values of s and r.

As a first step, we apply conventional IEN methods, based on assumed asymptotic normality of the Z-scores, to obtain a crude estimate of σ_α^2. While these conventional IEN methods are not suitable for flagging and hypothesis testing purposes, we observe empirically in Section 7.1 that the crude estimate $\hat{\sigma}_\alpha^2$ will often be accurate enough to aid in defining suitable initial values for s and r. In cases where $\hat{\sigma}_\alpha^2$ is far from the true value, a suitable optimization algorithm would still be able to maximize the likelihood, and would be no worse than an implementation with arbitrary initial values. Therefore, the asymptotic IEN methods are used in this step only to make a potentially informative first guess at the true parameter values, and they do not impact the performance of our proposed Exact IEN method.

It is often assumed in IEN methodology that α_i has a mean of zero, conditional on the observed covariates. This is a very mild assumption, because in situations where this does not hold, α_i can be redefined to satisfy the condition by adopting a Correlated Random Effects model framework.[12] Under our proposed model, the main assumption is that $\lambda_i = \exp(\alpha_i) \sim \text{Gamma}(s, r)$, and combining this with the mild assumption on the conditional moments of α_i, we have

$$E[\log(\lambda_i)] = E[\alpha_i] = \psi(s) - \log(r) = 0,$$

$$\text{Var}[\log(\lambda_i)] = \text{Var}[\alpha_i] = \psi_1(s) = \sigma_\alpha^2,$$

where $\psi(\cdot)$ is the digamma function and $\psi_1(\cdot)$ is the trigamma function.[20] By plugging in $\hat{\sigma}_\alpha^2$ from the first step, we obtain the initial values $s^{(0)} = \psi_1^{-1}(\hat{\sigma}_\alpha^2)$ and $r^{(0)} = \exp(\psi(s^{(0)}))$.

4.4 | Funnel plots

Funnel plots are visual representations of healthcare provider performance, where quality measure ratios (ie, O_i/E_i) are plotted against effective provider size, with control limits to flag providers that significantly deviate from the national expectations.[4] For flagging approaches that are based on test statistics with continuous null distributions, such as the Z-scores in (2), the control limits can be derived for a certain significance level by inverting these statistics onto the measure ratio scale.[7] On the other hand, for flagging approaches that are based on exact hypothesis tests with discrete null distributions, such as the Naive Exact Poisson test, authors have proposed interpolation methods to compute appropriate control limit thresholds.[21] We may directly apply these interpolation methods to construct funnel plots for our proposed IEN method because it is based on a discrete Negative Binomial null distribution. Examples of this implementation are provided in Sections 8.1 and 8.2.

4.5 | Illustrative numerical example

We now present a simple example on a simulated dataset to visually demonstrate the limitations of asymptotic IEN methods and the motivation for our proposed framework. We set $I = 1000$ and $\sigma_\alpha^2 = 0.5$, and for illustration purposes, we let $\tilde{n}_i = E_i = 10$ for all i and simulate λ_i from a Gamma($s = 2.5, r = 2$) distribution. In this example, we use a constant $\tilde{n}_i$ so that we may easily visualize the true and estimated empirical null distributions, but these methods are designed to also handle situations with varying $\tilde{n}_i$. For 90% of providers, γ_i^* is set to zero to simulate average care, and for the other 10%, γ_i^* is set to one to simulate low-quality care. Then, for each provider, O_i is simulated from a Poisson($E_i\lambda_i \exp(\gamma_i^*)$) distribution. We estimate σ_α^2 from asymptotic IEN methods, and s and r from our proposed exact IEN methods.

Figure 1 shows the marginal empirical distributions of the test statistics, from null providers with $\gamma_i^* = 0$, and the estimated null distributions overlaid. In this example, the effective provider sizes are small and unobserved confounding is present. As a result, we observe that the marginal empirical distribution of the null Z-scores is highly skewed and lacks any resemblance to the normal density. Thus, we see that if asymptotic IEN methods are applied to this simulated dataset, the estimated null distribution provides a very poor fit to the null Z-scores. For example, the lower tail of the estimated null distribution is too heavy, which would reduce our ability to detect providers with extreme quality of care effects in that direction. In contrast, the estimated Negative Binomial distribution from our proposed Exact IEN method fits the null observed outcome counts well. Detailed simulation studies are provided in Section 7 to assess the statistical properties of our proposed method in comparison with conventional IEN methods.

5 | INDIVIDUALIZED EMPIRICAL NULL FOR EXACT POISSON-BINOMIAL TESTS

5.1 | Framework

In some applications, the outcome of interest will be recorded as a binary random variable. For example, in Section 8.2, we present an analysis of transplant centers, where the outcome is related to whether the center accepts or declines an offered organ for transplantation. In this section, we show that the framework from Section 4.1 can be extended to settings with Bernoulli outcomes that are relatively rare.

If $Y_{ij}|\theta_{ij}^*$ follows a Bernoulli distribution with probability $b'(\theta_{ij}^*) = \exp(\theta_{ij}^*)/(1+\exp(\theta_{ij}^*))$, then O_i follows a Poisson-Binomial distribution under the null hypothesis (conditional on α_i), with $E[O_i|\gamma_i^* = 0, \alpha_i] = \sum_{j=1}^{n_i} \exp(\alpha_i + \theta_{ij}^0)/(1+\exp(\alpha_i+\theta_{ij}^0))$. By Le Cam's theorem,[22] the Poisson-Binomial null distribution is closely approximated by a Poisson null distribution with the same mean. More specifically, if we denote $m_i = E[O_i|\gamma_i^* = 0, \alpha_i]$ and $\wedge$ as the minimum, then the total variation distance between the Poisson-Binomial and Poisson null distributions has the following upper bound:

$$\sum_{c=0}^{\infty}\left|P(O_i = c|\gamma_i^* = 0, \alpha_i) - \frac{m_i^c e^{-m_i}}{c!}\right| \le 2\left(1 \wedge \frac{1}{m_i}\right)\sum_{j=1}^{n_i}\left\{\exp\left(\alpha_i + \theta_{ij}^0\right) \Big/ \left(1+\exp\left(\alpha_i+\theta_{ij}^0\right)\right)\right\}^2,$$

which is very small for rare outcomes. Thus, the framework for Poisson outcomes in Section 4.1 can be adapted to provide a close approximation for the exact null distribution. We argue that exact hypothesis tests are especially needed in

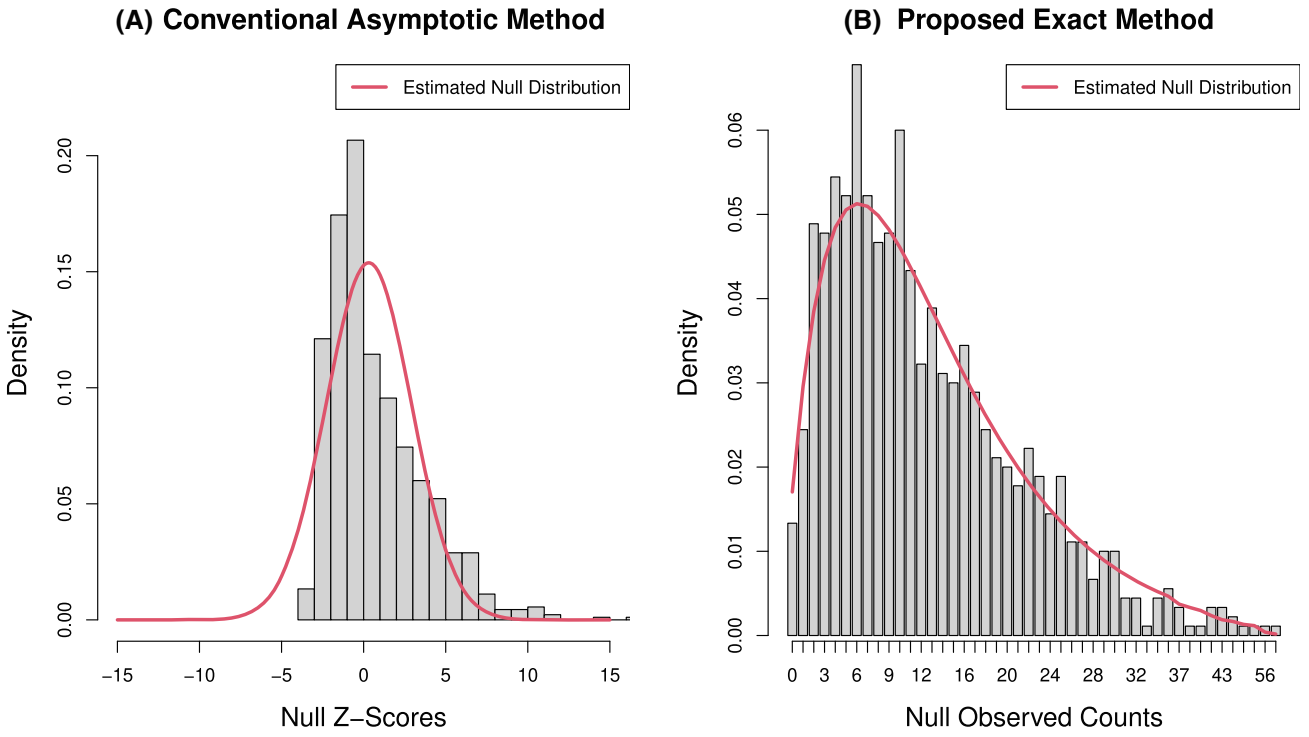


FIGURE 1 (A) Histogram of naive Z-scores for simulated null providers delivering average care, with the estimated empirical null (EN) distribution overlaid, based on conventional asymptotic IEN methods. (B) Barplot of observed outcome counts for simulated null providers delivering average care, with the estimated EN distribution overlaid, based on the proposed Exact IEN method. Results are based on 900 simulated null centers, each with a small effective provider size of $\tilde{n}_i = 10$. The variance of the unobserved confounding term is set to $\sigma_\alpha^2 = 0.5$. The exact EN distribution provides a better fit to the null counts than the asymptotic EN distribution does to the null Z-scores.

applications with rare outcomes, where the effective provider sizes are likely small, and our proposed approximation is most accurate under these settings. Furthermore, while our theoretical arguments require the assumption that the Bernoulli outcome is relatively rare, we will show in Sections 7.3 and 8.2 that the proposed method is fairly robust to this assumption and useful across a wide range of scenarios.

5.2 | Initial values

For Bernoulli outcomes, the canonical link function of the underlying GLM is different from that of an underlying Poisson model. Therefore, λ_i no longer equals $\exp(\alpha_i)$ as we had assumed for the Poisson model in Section 4.3, and even though the exact null distribution of O_i may be well-approximated by a Poisson distribution, we face new challenges in identifying appropriate initial values for the parameters in our optimization routine. In this section, we propose initial $s^{(0)}$ and $r^{(0)}$ values that we argue are reasonably close to the true parameter values.

Taking a first-order Taylor series expansion of m_i around θ_{ij}^0, we have

$$m_i \approx \sum_{j=1}^{n_i} \left[\frac{\exp(\theta_{ij}^0)}{1+\exp(\theta_{ij}^0)} + \frac{\exp(\theta_{ij}^0)}{(1+\exp(\theta_{ij}^0))^2} \alpha_i \right].$$

Then, following the framework in Section 4.1, we assume that $m_i \sim \text{Gamma}(s, r_i)$, where the rate r_i potentially depends on $\tilde{n}_i$. Combining this assumption with the approximation above,

$$E[m_i] = \frac{s}{r_i} \approx \sum_{j=1}^{n_i} \frac{\exp(\theta_{ij}^0)}{1+\exp(\theta_{ij}^0)} \approx \sum_{j=1}^{n_i} \frac{\exp(\theta_{ij}^0)}{(1+\exp(\theta_{ij}^0))^2},$$

where the second approximation is again a result of Le Cam's Theorem, and

$$\text{Var}[m_i] = \frac{s}{r_i^2} \approx \sigma_\alpha^2 \left[\sum_{j=1}^{n_i} \frac{\exp(\theta_{ij}^0)}{(1+\exp(\theta_{ij}^0))^2} \right]^2.$$

Plugging in the crude estimate, $\hat{\sigma}_\alpha^2$, from asymptotic IEN methods,

$$s^{(0)} = \frac{1}{\hat{\sigma}_\alpha^2},$$
$$r_i^{(0)} = \frac{r^{(0)}}{\sum_{j=1}^{n_i} \frac{\exp(\theta_{ij}^0)}{(1+\exp(\theta_{ij}^0))^2}} = \frac{r^{(0)}}{\tilde{n}_i}, \text{ where } r^{(0)} = \frac{1}{\hat{\sigma}_\alpha^2}.$$

5.3 | Estimation

From the results above in Sections 5.1 and 5.2, we assume that O_i approximately follows a Poisson(m_i) distribution under the null hypothesis (conditional on α_i), with $m_i \sim \text{Gamma}(s, r/\tilde{n}_i)$. This yields the following marginal null distribution:

$$(O_i | \gamma_i^* = 0) \sim \text{NB}\left(s, \frac{r/\tilde{n}_i}{1 + r/\tilde{n}_i}\right), \tag{6}$$

where $\tilde{n}_i = \sum_{j=1}^{n_i} \frac{\exp(\theta_{ij}^0)}{(1+\exp(\theta_{ij}^0))^2}$. Recall that in Equation (4) from Section 4.1, $E_i = \tilde{n}_i$, which implies that the derived null distribution in Equation (6) for Bernoulli outcome settings is analogous to our proposed null distribution for Poisson outcome settings. To estimate s and r, we use the initial values from Section 5.2 and apply our proposed optimization routine, as described in Section 4.3.

6 | PATIENT-LEVEL UNOBSERVED CONFOUNDING

Throughout this article, we have developed methods to adjust for a provider-level unobserved random quantity, α_i, which has the potential to bias the estimates and inference of the provider-level healthcare quality parameter γ_i^*. This α_i term may reflect the combined influence of many unmeasured factors that are beyond the providers' control, affect patient outcomes, and vary in severity across the providers. As introduced in Section 3, some examples of these unmeasured provider-level factors may include the latent impact of the COVID-19 pandemic on patient communities, or disparities in the amount of resources allocated to the providers. Unobserved confounding may also arise due to patient-level risk factors that are unmeasured and excluded from the $\boldsymbol{X}_{ij}$ term in Equation (1). In this section, we show formally that our proposed IEN framework is also able to account for the between-provider variation in these unmeasured patient-level factors under certain conditions.

First, assume that an unobserved patient-level quantity α_{ij} can be written as a linear combination of K unmeasured patient covariates $(W_{ij1}, \ldots, W_{ijK})$:

$$\alpha_{ij} = \sum_{k=1}^{K} \xi_k W_{ijk}.$$

Then, we may decompose α_{ij} into between-provider and within-provider terms:

$$\alpha_{ij} = \sum_{k=1}^{K} \xi_k \overline{W}_{ik} + \sum_{k=1}^{K} \xi_k \left(W_{ijk} - \overline{W}_{ik} \right),$$

where $\overline{W}_{ik} = \frac{1}{n_i} \sum_{j=1}^{n_i} W_{ijk}$ is the provider-level sample mean of the covariate. We denote the first term as $\alpha_i = \sum_{k=1}^{K} \xi_k \overline{W}_{ik}$, and we may account for α_i through the proposed IEN adjustments. The second term $\varepsilon_{ij} = \sum_{k=1}^{K} \xi_k (W_{ijk} - \overline{W}_{ik})$ is a residual within-provider variable that does not directly bias the provider-level treatment effect estimates, but can lead to

model misspecification by potentially biasing the $\boldsymbol{\beta}$ coefficient vector of the observed patient-level characteristics $\boldsymbol{X}_{ij}$ (see Equation (1) from Section 3). However, it has been shown that ϵ_{ij} does not introduce bias if it is independent of $\boldsymbol{X}_{ij}$ and the canonical link function for Model (1) is the identity or log link function,[23] or if it is conditionally independent of $\boldsymbol{X}_{ij}$, given Y_{ij}, and the canonical link function for Model (1) is the logit link function.[24] Thus, in provider profiling applications where these three link functions are by far the most widely used in practice, the impact of an independent (or conditionally independent) ϵ_{ij} can be ignored.

7 | SIMULATIONS

7.1 | Parameter estimation

In this section, we assess the accuracy of our proposed IEN estimation methods through simulations. For simplicity, we focus on the setting in which Y_{ij} follows a Poisson distribution. For $I = 100$ providers, we set $n_i = 100$ and simulate α_i from a Log-Gamma distribution with mean zero and variance σ_α^2, where σ_α^2 is varied from zero to one. This implies that $\lambda_i = \exp(\alpha_i)$ follows a Gamma(s, r) distribution, where the parameters s and r change with σ_α^2. In Section 7.3, we consider scenarios in which this Gamma distributional assumption does not hold, and we show that our proposed method stays robust to model misspecification. The $E_i = \tilde{n}_i, i = 1, \ldots, I$ are simulated from a Uniform(1,100) distribution such that some providers have small effective sizes. For 90 providers, the quality of care effects, γ_i^*, are set to zero to simulate average care, and for the remaining 10 providers, $\gamma_i^* = \pm 1$ to simulate care of extremely low or high quality. The observed number of outcomes for the ith provider, O_i, is simulated from a Poisson distribution with mean $E_i \exp(\gamma_i^*)\lambda_i$.

We evaluate the performances of conventional IEN methods in estimating the σ_α^2 parameter, which is used to define a normal null distribution, and compare this to the performance of our proposed Exact IEN method in estimating the s and r parameters, which are used to define a Negative Binomial null distribution. Figure 2 shows that the $\hat{s}$ and $\hat{r}$

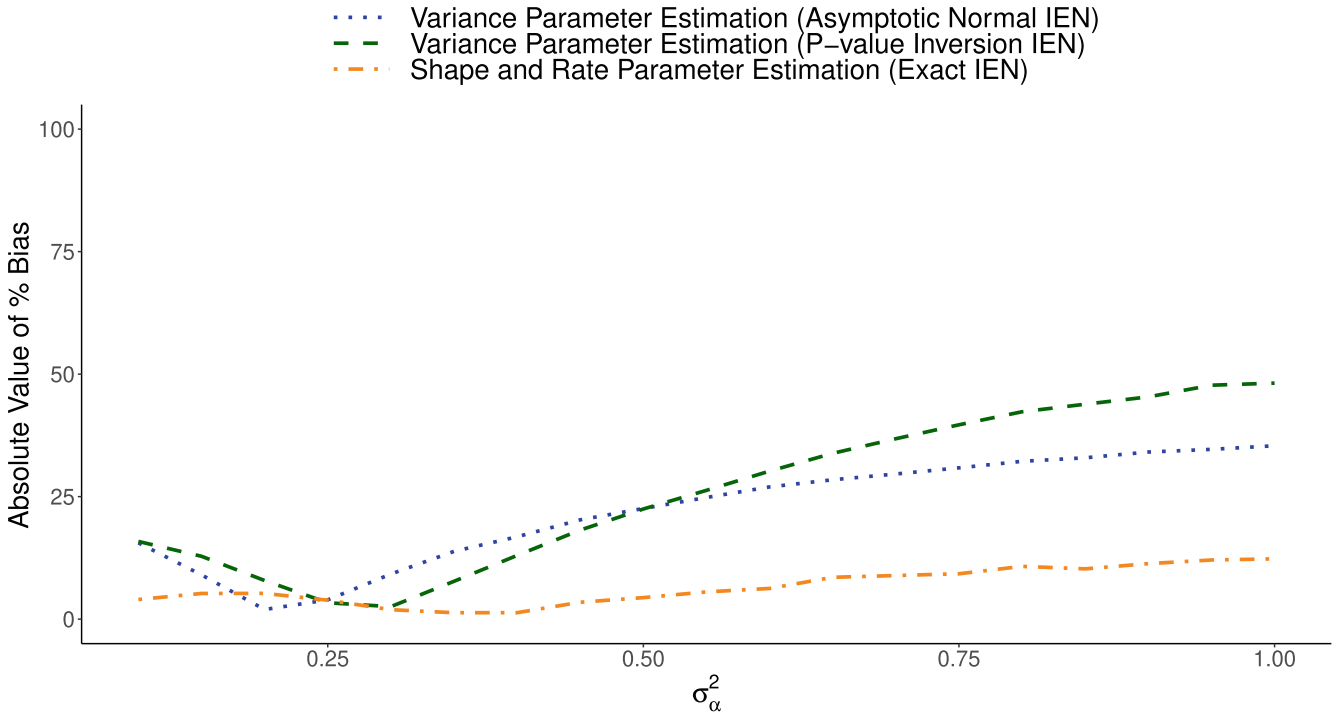


FIGURE 2 Absolute value of the percent bias for different individualized empirical null (IEN) estimation strategies, and varying levels of unobserved confounding. The severity of the unobserved confounding increases with σ_α^2. The conventional Asymptotic Normal IEN and *P*-value Inversion methods estimate a null variance parameter (σ_α^2) to define normal densities for the null model. The proposed Exact IEN method estimates null shape and rate parameters (which depend on σ_α^2) to define Negative Binomial densities for the null model. The dot-dashed line shows the average absolute value of the percent bias for the shape and rate parameters. The simulation results are based on 5000 iterations.

estimates from the proposed Exact IEN method are less biased than the $\hat{\sigma}_\alpha^2$ estimates from traditional IEN methods. While all of the estimates become more biased as σ_α^2 (the variance of the unobserved confounding term) increases, this is less severe for $\hat{s}$ and $\hat{r}$ than it is for the $\hat{\sigma}_\alpha^2$ from conventional IEN methods. The main limitation of the conventional IEN methods is that they rely on the asymptotic normal density to specify the null distribution, which is often an inappropriate model and can inhibit the accurate detection of providers with low-quality care, as we will show in Section 7.2. However, these simulations uncover the additional limitation that the null parameter estimation is also impacted by the use of an asymptotic normal model, and the estimated σ_α^2 for the asymptotic normal null distribution is more biased than the s and r estimates for our proposed Negative Binomial null distribution (Figure 2).

We find through our evaluations that the s and r parameter estimates from our proposed method may occasionally be highly extreme when σ_α^2 is very close to zero. Intuitively, this happens because there is almost no overdispersion when σ_α^2 is small, and larger s and r values can increase the likelihood by making the variance of the proposed Negative Binomial null density similar to the mean, as in the naive Poisson null distribution that assumes $\sigma_\alpha^2 = 0$. Therefore, even in cases where these extreme s and r estimates are far from the true parameter values, they may correctly indicate that overdispersion due to unobserved confounding is only a very minor issue, and the flagging results from IEN methods would be similar to those from naive methods anyway. In Figure 2, we do not include a simulation setting with σ_α^2 very close to zero, as this setting is not realistic or of interest (ie, there is almost no overdispersion to correct for) and the scale of the absolute bias axis would become highly distorted and reduce the visual clarity.

7.2 | Flagging rates and coverage probability

Using the simulation settings from Section 7.1, we compare the rates of false and true flagging across different IEN methods (using thresholds from a 5% significance level). In Figure 3, we observe that the Naive Exact test, which does not make any IEN correction for unobserved confounding, has a very high rate of false flagging, which becomes more severe

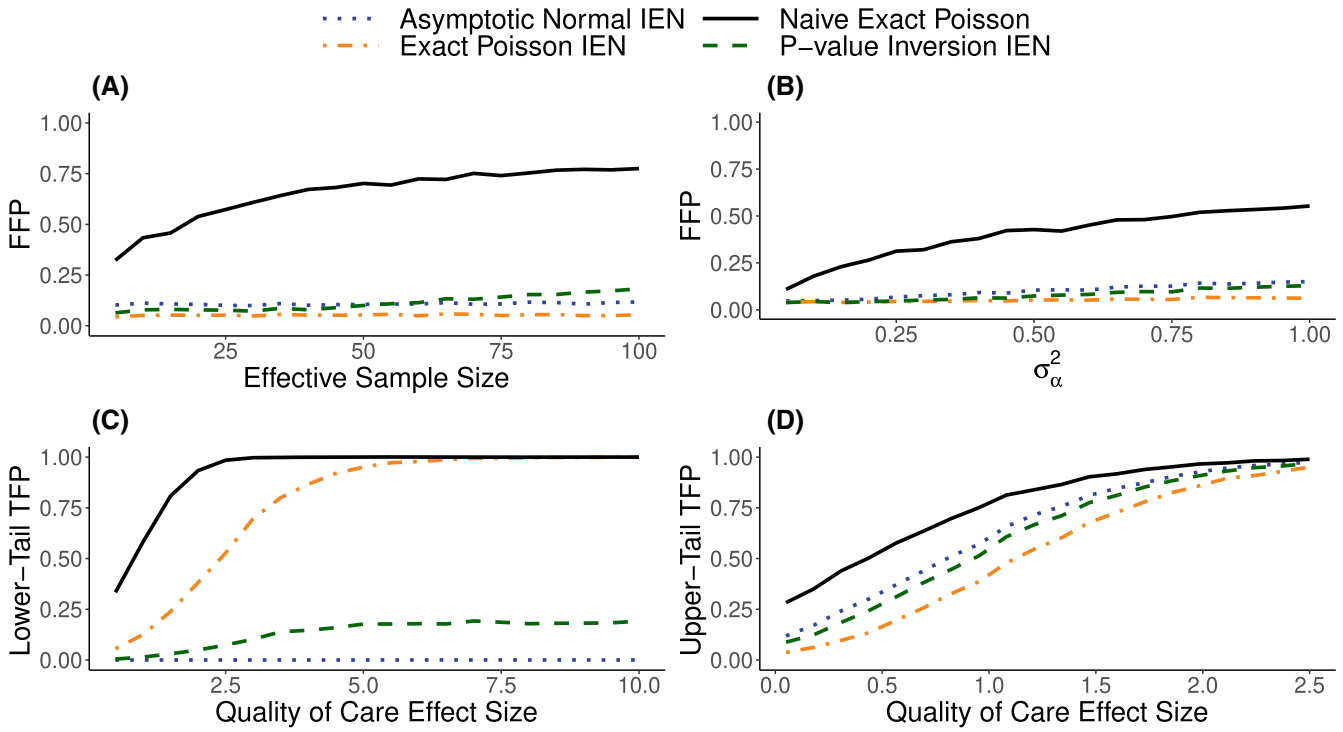


FIGURE 3 False-flagging probabilities (FFP) and true-flagging probabilities (TFP) for different individualized empirical null (IEN) estimation methods. The outcome data are generated from Poisson distributions. (A) FFP by effective provider size ($\tilde{n}_i$). (B) FFP by the variance of the unobserved confounding term (σ_α^2). (C) TFP in the lower-tail of the null distribution, by the quality of care effect size ($|\gamma_i^*|$). (D) TFP in the upper-tail of the null distribution, by the quality of care effect size ($|\gamma_i^*|$). The simulation results are based on 5000 iterations. The Naive Exact test does not correct for overdispersion due to unobserved confounding, the Asymptotic Normal and *P*-value Inversion IEN methods are existing corrections, and the Exact Poisson IEN is the proposed correction. The simulation results are based on 5000 iterations.

as $\tilde{n}_i$ increases. As discussed in Section 3, this behavior is well-documented in the provider profiling literature, and it is a consequence of overdispersion in the test statistics that is caused by unobserved confounding. The conventional IEN methods, which attempt to correct for this overdispersion and its strong impact on larger providers, still have elevated false-flagging probabilities (FFPs). In addition, the *P*-value Inversion IEN method shows a minor sensitivity to the effective provider size, where the FFP increases with $\tilde{n}_i$. These results provide evidence that the conventional IEN models are inappropriate for settings where small providers are under evaluation or the risk-adjustment model is misspecified, and even if the provider size is large, our simulation results suggest that the model misspecification itself can cause the asymptotic normality assumption to be violated. In contrast, the proposed IEN method for Exact Poisson tests has an FFP that is very close to zero, regardless of the magnitude of $\tilde{n}_i$ or σ_α^2 (Figure 3).

Figure 3 also shows that existing IEN methods have severe difficulty in detecting non-null providers with test statistics in the lower-tails of the null distributions. The conventional IEN methods (based on asymptotic normality or *p*-value inversion) have extremely low power to detect providers with negative coefficients for the quality of care effects, and this power does not improve, even as the effect sizes, $|\gamma_i^*|$, become extremely large. On the other hand, the proposed exact IEN method has a lower-tail true flagging probability (TFP) that increases rapidly with $|\gamma_i^*|$ and is much higher than that of conventional methods. Based on the simulated example in Figure 1, we suspect that these results reflect the poor fit of the symmetric normal distribution that is used in conventional IEN methods. For applications with small providers or misspecified models, the true null distribution may be highly skewed, and the asymptotic normal density would have one tail that is too heavy and one tail that is too light, reducing TFPs in one direction while increasing FFPs. The proposed exact IEN test is the only method that keeps the FFPs low while maintaining high power to detect unusual healthcare quality in either effect direction.

For each of the inference methods compared above, we compute the corresponding 95% confidence interval (CI) for the true value of $\exp(\gamma_i^*)$ and perform additional simulations to assess the average lengths and coverage probabilities of these CIs. We observe in Supplementary Figure 3 that the CI for the proposed Exact IEN method has a coverage probability that is very close to the nominal level of 0.95, regardless of the value of γ_i^*. In contrast, all existing inference methods suffer from a very large drop in coverage probability as γ_i^* increases. Our comparisons of the average CI length also reflect the flagging results from Figure 3. For negative γ_i^* values that correspond to flagging in the lower tail of the null distribution, the Exact IEN method has more power than existing IEN methods and also a narrower CI. In contrast, for γ_i^* values that correspond to flagging in the upper tail of the null distribution, the Exact IEN method appropriately has a wider CI than the existing IEN methods, which tend to be less accurate under these settings (Figure 3 and Supplementary Figure 3).

7.3 | Model robustness

While our proposed IEN method does not require any assumptions regarding the asymptotic properties of the test statistics, it does rely on certain distributional assumptions to define the null model. For example, the proposed IEN method for Poisson outcomes assumes that λ_i follows a Gamma(s, r) distribution. Also, in Section 5.1, we argue that the Negative Binomial null distribution from Section 4.1 can be adapted to the Bernoulli outcome setting when the outcome is rare enough. In this section, we assess the robustness of our method and study the flagging performance under scenarios where these conditions are violated.

To evaluate the impact of violations in the Gamma distributional assumption, we simulate the observed outcome counts, O_i, from a Poisson distribution, as in Section 7.1, and allow λ_i to be generated from distributions other than the assumed Gamma(s, r) density. We consider settings where λ_i follows a Lognormal$(0, \sigma_\alpha^2)$ or Weibull$(\pi/(6\sigma_\alpha), \exp(-6\sigma_\alpha\psi(1)/\pi))$ distribution, and we compare them to the setting where the underlying model matches the assumed model (ie, $\lambda_i \sim$ Gamma(s, r)). Each setting is designed such that $E[\log(\lambda_i)] = E[\alpha_i] = 0$ and $\text{Var}[\log(\lambda_i)] = \text{Var}[\alpha_i] = \sigma_\alpha^2$. Supplementary Figure 4 shows that the underlying distribution of λ_i has little impact on the performance of the proposed IEN method. Even when the distribution of λ_i is misspecified, the FFPs of the proposed IEN method remain very low and are similar to those from the correctly-specified null model.

We also consider the scenario in which O_i does not follow a Poisson distribution as originally assumed in the Exact Poisson test. In Section 4.1, we claimed that the proposed IEN method would still control FFPs if O_i followed a distribution that has a variance larger than its mean. Conceptually, this is because the IEN method is designed to account for overdispersion due to residual unobserved confounding, and any additional overdispersion from the original model would be absorbed into the estimated null distribution. We assess this claim through simulation by generating $O_i|\alpha_i, \gamma_i^* = 0$

from a Negative Binomial distribution with mean $E_i\lambda_i$ and variance $E_i\lambda_i(1 + 2E_i\lambda_i)$, instead of the assumed Poisson$(E_i\lambda_i)$ distribution. Supplementary Figure 5 shows that the FFPs for the Naive Exact Poisson test increases after introducing this additional overdispersion. However, the FFPs for all three IEN methods are stable, even when the original Poisson distributional assumption is violated.

The proposed IEN approximation for Bernoulli outcomes, described in Section 5.1, departs from the true null model as the patient outcomes become more frequently observed, and the most severe consequence of this departure is the potential for less accurate flagging. Thus, to assess the robustness of our proposed IEN method, we simulate binary outcome data with varying levels of overall outcome frequency and compute the FFPs and TFPs numerically. On each iteration, we generate α_i from a Log-Gamma distribution with mean zero and variance σ_α^2, and we generate $n_i = 100$ realizations of an observed patient-level covariate X_{ij}, from a $N(m_{X,i}, 1)$ distribution, where $m_{X,i} \sim N(0, 1)$. The quality of care effects, γ_i^*, are specified as in Section 7.1. The population norm, μ, is varied to achieve different national outcome rates, and O_i is generated from a Poisson-Binomial distribution with expected probability vector $(ep_{i1} \ \dots \ ep_{in_i})^\top$, where $ep_{ij} = \exp(\mu + \gamma_i^* + \alpha_i + X_{ij})/\{1 + \exp(\mu + \gamma_i^* + \alpha_i + X_{ij})\}$.

In Supplementary Figure 6, we observe that the proposed method has a lower FFP and higher TFP, compared to existing IEN methods, when the national outcome proportion is less than approximately 0.35. As the national outcome proportion approaches 0.5, making our proposed approximation less accurate and the $\tilde{n}_i$ values larger, the conventional IEN methods have a very slight advantage over the exact IEN method in terms of FFPs and TFPs. These results show that the proposed method is fairly robust, even if the outcome is not extremely rare, and in many practical situations, the proposed method may have superior overall accuracy compared to conventional IEN approaches. Furthermore, if the outcome is very frequently observed (eg, the national proportion exceeds 0.5), one may use $1 - Y_{ij}$ as the outcome variable to achieve more accurate evaluations. For healthcare quality measures that are based on binary outcomes with a national outcome proportion near 0.5 (eg, approximately 0.35 to 0.65 from our simulations), conventional IEN methods may be slightly preferred.

7.4 | Computational costs

To compare the computational costs of the existing and proposed inference methods, we measure the speed and memory allocation of the estimation and flagging procedures on 5000 simulated datasets, using the same settings as in Section 7.2. In Supplementary Table 1, we observe that all of the methods run very quickly with little memory usage, so the computational cost is not a barrier to implementing any of these methods. While our proposed Exact IEN method is slightly more computational expensive than existing methods, it still completes the null distribution estimation and flagging for all providers in less than one second with very manageable memory usage, so these differences are negligible and would likely not be noticed by users.

8 | REAL DATA ANALYSES

8.1 | Standardized mortality ratio (SMR)

We apply the proposed IEN methods for exact Poisson tests (Section 4.1) to assess mortality outcomes within dialysis facilities. The facility-level data used in this analysis are publicly available through the U.S. Dialysis Facility Reports, which are published by the CMS to inform patient care.[25] From these public reports, we obtain the observed and expected numbers of deaths, based on three-year data (years 2017-2019) for the 7018 U.S. dialysis facilities that are evaluated by the CMS.

The Exact Poisson test is a very popular method for assessing providers with respect to the SMR.[16] In most applications, this is because there are small healthcare facilities that treat few patients and have low expected mortality rates, so the effective provider sizes are close to zero and assumptions about the asymptotic properties of the test statistics may not hold. However, IEN methods have not been available for the Exact Poisson test, and because of this, the *P*-value Inversion method described in Section 3 is currently used in practice to identify U.S. dialysis facilities that provide low-quality care in terms of the SMR measure.[15]

In Figure 4, we show the empirical marginal distribution of the naive *Z*-scores for 320 dialysis facilities with very similar effective provider sizes ($1 < \tilde{n}_i < 2$) and statistics that fall within the null intervals. If the large-sample arguments

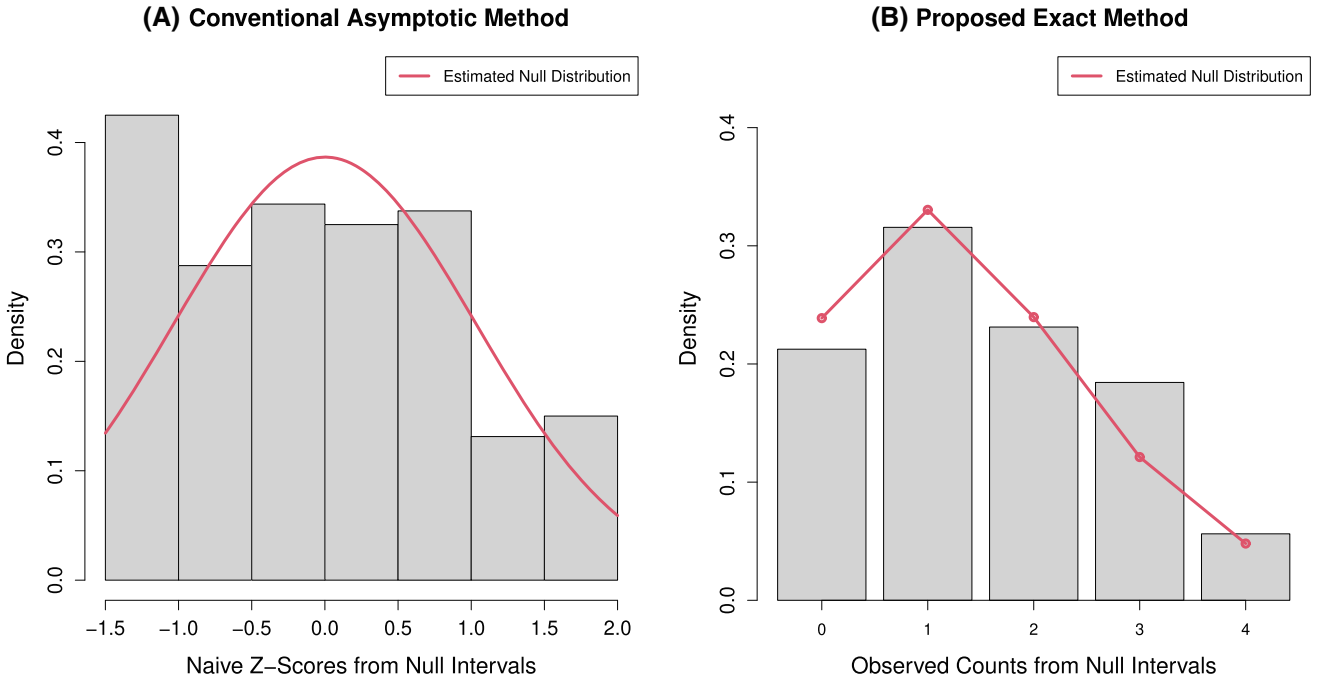


FIGURE 4 (A) Histogram of naive Z-scores for the Standardized Mortality Ratio (SMR), from 320 dialysis facilities with small effective provider sizes ($\tilde{n}_i$) between one and two, and statistics that fall within the assumed null intervals. The estimated empirical null distribution is overlaid (approximated with $\tilde{n}_i = 1.5$ for all of the facilities), based on conventional asymptotic IEN methods. (B) Barplot of the observed death counts, from 320 dialysis facilities with small effective provider sizes between one and two, and statistics that fall within the assumed null intervals. The estimated empirical null distribution is overlaid (approximated with $\tilde{n}_i = 1.5$ for all of the facilities), based on the proposed exact Poisson IEN method. The effective provider size is defined here as the expected number of deaths, conditional on a set of observed patient-level confounding variables.

of existing IEN methods were to hold, we would expect this empirical distribution to be symmetric and closely resemble a normal null density, since $\tilde{n}_i$ is nearly the same across all facilities and we suspect that these facilities provide average care. However, we observe instead that the empirical distribution is highly skewed and the normal density provides a poor fit to the data, similar to the simulated example in Figure 1. Thus, we argue that for some facilities in our dataset, such as the ones in Figure 4, the asymptotic normal density is not a suitable model for the null Z-scores, and exact tests should be used.

We compare the flagging results for all 7018 facilities based on four different testing methods (using thresholds from a 5% significance level): the Naive Exact test without any IEN corrections, the Asymptotic Normal IEN test, the P-value Inversion IEN test, and our proposed Exact IEN test. Table 1 shows that the Naive Exact Poisson test (without any IEN correction) flags a large proportion of facilities for deviating substantially from national quality standards, whereas all three IEN methods flag substantially fewer facilities (Table 1 and Supplementary Table 2). Based on the simulation results in Section 7, it is reasonable to assume that the elevated flagging rate of the Naive Exact Poisson test is driven by unobserved confounding, and many of the flagged facilities from this method are not actually deviating from the national norms. We also observe that our proposed Exact Poisson IEN method flags a larger number of facilities for high-quality care (ie, low SMR values), compared to the Asymptotic Normal IEN method and the P-value Inversion IEN method conventionally used in practice. On the other hand, the Exact Poisson IEN method flags fewer facilities for low-quality care (ie, high SMR values), compared to the conventional IEN methods. These results are consistent with the simulation results in Section 7, which showed that the proposed IEN method has substantially more power to detect true departures from the national quality standards in the lower-tail of the null distribution, compared to conventional IEN methods, and it reduces false flagging rates that are mainly driven by statistics that fall in the upper tail of the null distribution.

The funnel plots in Supplementary Figure 7 provide further insight into the advantages of the proposed Exact Poisson IEN method. As expected, the control limits from the Naive Exact Poisson test are generally much closer to the null ratio of one, compared to the IEN methods, causing this method to flag many facilities (Table 1). For large facilities, the control limits of all three IEN methods are similar to each other. However, for very small facilities, the Asymptotic Normal IEN

TABLE 1 Counts (%) of U.S. dialysis facilities that are labeled as providing care of low, average, or high quality, in terms of Standardized Mortality Ratios (SMRs) and based on various flagging approaches.

	Quality of care		
Testing method	**Low**	**Average**	**High**
Naive exact	399 (5.7)	6318 (90.0)	301 (4.3)
Asymptotic normal IEN	355 (5.1)	6610 (94.2)	53 (0.8)
P-value inversion IEN	250 (3.6)	6639 (94.6)	129 (1.8)
Exact IEN	240 (3.4)	6645 (94.7)	133 (1.9)

Note: Results are shown for 7018 total facilities that are evaluated by the Centers for Medicare and Medicaid Services. The Naive Exact method does not account for overdispersion due to unobserved confounding, whereas the IEN methods attempt to correct for this. The Asymptotic Normal and *P*-value Inversion IEN methods are existing corrections, and the Exact IEN is the proposed correction. IEN: Individualized Empirical Null.

method has the most liberal control limits in the upper tail of the SMR distribution, with very conservative control limits in the lower tail. This provides evidence that the asymptotic normality assumption is violated for these small facilities, and the assumed symmetric null distribution causes overly-liberal flagging in one direction and overly-conservative flagging in the other direction (Supplementary Figure 7), as seen in our simulated examples (Sections 4.5 and 7.2). In this application, the *P*-value Inversion method has control limits that are relatively similar to our proposed Exact Poisson IEN method; we show in Section 8.2 that this is not always the case.

8.2 | Standardized acceptance ratio (SAR)

To demonstrate the application of our proposed Exact IEN methods for Bernoulli outcomes, we evaluate the performances of kidney transplant centers in accepting donated organs. End-stage renal disease (ESRD) is a severe illness, and it has been shown that transplantation is the best form of treatment for most ESRD patients.[26] However, U.S. ESRD patients face long wait times to receive kidney transplants, and despite the wide and growing demand for donated organs, many of these organs are wasted.[27,28] In response to this inefficient use of resources, the Scientific Registry of Transplant Recipients (SRTR) and the CMS have led efforts to monitor the frequency at which transplant centers accept offered organs.[1]

The SAR is a healthcare quality measure that is used to assess offer acceptance practices and identify unusually low or high acceptance rates.[29] It compares a center's observed number of accepted organs to the expected number given national standards and patient risk adjustment. In this application, the outcome data are represented by a binary random variable Y_{ij}, indicating whether the center accepted an organ that was offered to it. The national organ offer acceptance rate in the U.S. is very low ($< 1\%$ of offers were accepted during the cohort time period of July 1, 2020 through June 20, 2021), so the outcome of interest is rare and we argue that our proposed null model approximation, as described in Section 5.1, is suitable for this setting.

We compare the flagging results for 232 kidney transplant programs, using public data from the SRTR's Program Specific Reports[29] and the following testing methods (with thresholds from a 5% significance level): the Naive Exact test without any IEN corrections, the Asymptotic Normal IEN test, the *P*-value Inversion IEN test, and our proposed Exact IEN test. As in Section 8.1, the Naive Exact test, without any IEN corrections, flags a much larger number of centers compared to all three IEN methods (Table 2 and Supplementary Table 3). In addition, our proposed IEN method for the Exact Poisson-Binomial test flags more centers for low-quality care (ie, low SAR values) and fewer centers for high-quality care (ie, high SAR values), compared to the Asymptotic Normal or *P*-value Inversion IEN methods (Table 2 and Supplementary Table 3). Again, these results are expected, as the simulations in Section 7 showed that the conventional IEN methods have difficulty detecting non-null providers with statistics in the lower tail of the null distribution, and they also have higher FFPs from statistics in the upper tail of the null distribution.

Supplementary Figure 8 shows the funnel plots that correspond to the flagging results described above. We observe that the Asymptotic IEN method has a control limit of zero for all effective provider sizes, which demonstrates that it would be impossible to ever detect a transplant center with low-quality care based on this approach. The *P*-value Inversion method and the proposed Exact Poisson IEN method have similar lower-tail control limits, but the upper-tail control limit of the *P*-value Inversion method is substantially closer to the null ratio of one, potentially reflecting the findings from our simulations that this method has a larger FFP compared to our proposed approach. The upper-tail control limit of the

TABLE 2 Counts (%) of U.S. kidney transplant programs that are labeled as providing care of low, average, or high quality, in terms of Standardized Acceptance Ratios (SAR) and based on various flagging approaches.

	Quality of care		
Testing method	**Low**	**Average**	**High**
Naive exact	75 (32.3)	81 (34.9)	76 (32.8)
Asymptotic normal, IEN	0 (0.0)	196 (84.5)	36 (15.5)
P-value inversion, IEN	2 (0.9)	199 (85.8)	31 (13.4)
Exact IEN	7 (3.0)	204 (87.9)	21 (9.1)

Note: Results are shown for the 232 programs that are included in the Scientific Registry of Transplant Recipients' public reports. The Naive Exact method does not account for overdispersion due to unobserved confounding, whereas the IEN methods do attempt to correct for this. The Asymptotic Normal and *P*-value Inversion IEN methods are existing corrections, and the Exact IEN is the proposed correction. IEN: Individualized Empirical Null.

P-value Inversion method also abruptly changes when the effective provider size is approximately equal to 50, and this is caused by the numerical floating point issues discussed in Section 3. With an effective provider size of approximately 50, the exact *p*-value needed to have a significant IEN-corrected pseudo-*Z*-score is so close to zero that the computer cannot store this number accurately. Thus, we may only replace this *p*-value with the smallest number that the software can handle, which may lead to inaccurate flagging results. These findings highlight the advantages of the proposed IEN method over the ad-hoc *P*-value Inversion method.

9 | DISCUSSION

We have proposed an IEN estimation approach for two popular exact tests of healthcare quality. These methods correct for the impact of unobserved confounding while recognizing the varying severity of this impact across provider size and avoiding assumptions about the asymptotic distributions of the evaluation metrics. The IEN approach for exact tests is particularly useful for applications that involve the evaluation of healthcare providers with small effective sizes and misspecified risk-adjustment models. Furthermore, our implementation of the proposed Exact IEN method has the same input and output data as the conventional inference methods that are currently used in practice for provider profiling. Thus, a transition to the Exact IEN framework would be practically convenient and support recent efforts to use novel inference methods for improved healthcare quality assessments.[5,7]

Unobserved confounding is a statistical problem that is challenging to address without imposing additional assumptions, and all existing corrections rely on various conditions to define a null model that incorporates the unobserved confounding. In this article, we have avoided all of the asymptotic assumptions of conventional IEN methods, which likely do not hold in provider profiling applications, at the cost of requiring one extra distributional assumption. Our simulations suggest that the proposed methods are much more robust to the assumptions of our proposed null model, as compared to conventional IEN methods, reducing FFPs and increasing TFPs across a wide range of settings. These advantages in flagging performance and robustness have important practical consequences, as government regulators use the flagging results from these tests to penalize providers for low-quality care and design interventions to improve the health system.

In many applications, statistical inference is conducted based on the asymptotic null distributions of the test statistics because it is mathematically convenient, and researchers can obtain larger samples by performing more data collection. However, in our motivating application of healthcare provider evaluations, the effective provider sizes are a function of how many patients the providers treat, which the investigator has no control over. In addition, the limitations of observational health services data make model misspecification a ubiquitous issue in the provider profiling field. Further research is needed to develop an even broader set of exact testing methods to evaluate healthcare providers fairly and account for the unavoidable impact of unobserved confounding. In a separate report, we further extend IEN methodology to allow for Bayesian statistical inference and incorporate observable provider-level confounding variables.[12]

AUTHOR CONTRIBUTIONS

NH and KH conceived of the article and reviewed and approved the final version. NH developed the statistical methods, performed the analyses, and wrote the manuscript draft.

FUNDING INFORMATION

This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases under Grant R01DK129539. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available from the SRTR and CMS public reports at https://www.srtr.org/reports/program-specific-reports/ and https://data.cms.gov/quality-of-care/medicare-dialysis-facilities.

ORCID

Nicholas Hartman https://orcid.org/0000-0002-8475-1973
Kevin He https://orcid.org/0000-0002-8354-426X

REFERENCES

1. Howard R, Cornell D, Schold J. CMS oversight, OPOs and transplant centers and the law of unintended consequences. *Clin Transpl*. 2009;23:778-783. doi:10.1111/j.1399-0012.2009.01157.x
2. Ash A, Fienberg S, Louis T, Normand S, Stukel T, Utts J. Statistical issues in assessing hospital performance. *Commissioned by the Committee of Presidents of Statistical Societies for the Centers for Medicare and Medicaid Services (CMS)*. Baltimore, MD: U.S. Centers for Medicare and Medicaid Services; 2012. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf
3. Kalbfleisch J, He K, Xia L, Li Y. Does the inter-unit reliability (IUR) measure reliability? *Health Serv Outcomes Res Methodol*. 2018;18:215-225. doi:10.1007/s10742-018-0185-4
4. Spiegelhalter D. Funnel plots for comparing institutional performance. *Stat Med*. 2005;24:1185-1202. doi:10.1002/sim.1970
5. He K, Kalbfleisch J, Li Y, Li Y. Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. *Lifetime Data Anal*. 2013;19:490-512. doi:10.1007/s10985-013-9264-6
6. Jones H, Spiegelhalter D. The identification of unusual health-care providers from a hierarchical model. *Am Stat*. 2011;65(3):154-163. doi:10.1198/tast.2011.10190
7. Spiegelhalter D, Sherlaw-Johnson C, Bardsley M, Blunt I, Wood C, Grigg O. Statistical methods for healthcare regulation: rating, screening and surveillance. *J R Stat Soc Ser A*. 2012;175(1):1-47. doi:10.1111/j.1467-985X.2011.01010.x
8. Xia L, He K, Li Y, Kalbfleisch J. Accounting for total variation and robustness in profiling health care providers. *Biostatistics*. 2022;23:257-273. doi:10.1093/biostatistics/kxaa024
9. Kalbfleisch J, Wolfe R. On monitoring outcomes of medical providers. *Stat Biosci*. 2013;5(2):286-302. doi:10.1007/s12561-013-9093-x
10. Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J Am Stat Assoc*. 2004;99(465):96-104. doi:10.1198/016214504000000089
11. Hartman N, Messana J, Kang J, Naik A, Shearon T, He K. Composite scores for transplant center evaluation: a new individualized empirical null method. *Ann Appl Stat*. 2024;18(1):729-748. doi:10.1214/23-AOAS1809
12. Hartman N, He K. Robust privacy-preserving models for cluster-level confounding: Recognizing disparities in access to transplantation. 2023. doi:10.48550/arXiv.2305.06208
13. Austin J, Kachalia A. The state of health care quality measurement in the era of COVID-19: the importance of doing better. *JAMA*. 2020;2020(4):333-334. doi:10.1001/jama.2020.11461
14. Efron B. Size, power and false discovery rates. *Ann Stat*. 2007;35(4):1351-1377. doi:10.1214/009053606000001460
15. University of Michigan Kidney Epidemiology and Cost Center. Standardized mortality ratio for dialysis facilities. 2022. https://dialysisdata.org/sites/default/files/content/ESRD_Measures/SMR_MIF.pdf
16. Liddell F. Simple exact analysis of the standardised mortality ratio. *J Epidemiol Community Health*. 1984;38:85-88. doi:10.1136/jech.38.1.85
17. Gauran I, Park J, Lim J, et al. Empirical null estimation using zero-inflated discrete mixture distributions and its application to protein domain data. *Biometrics*. 2018;74:458-471. doi:10.1111/biom.12779
18. Kalbfleisch J, Prentice R. *The Statistical Analysis of Failure Time Data*. 2nd ed. New York, NY: Wiley; 2002.
19. Nelder J, Mead R. A simplex method for function minimization. *Comput J*. 1965;7:308-313. doi:10.1093/comjnl/8.1.27
20. Halliwell L. The log-gamma distribution and non-normal error. *Variance*. 2018;13(2):173-189.
21. Manktelow B, Seaton S. Specifying the probability characteristics of funnel plot control limits: an investigation of three approaches. *PLoS One*. 2012;7(9):e45723. doi:10.1371/journal.pone.0045723
22. Le Cam L. An approximation theorem for the Poisson binomial distribution. *Pac J Math*. 1960;10:1181-1197.
23. Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*. 1993;80(4):807-815. doi:10.2307/2336872

24. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Stat Rev*. 1991;59(2):227-240. doi:10.2307/1403444
25. Centers for Medicare and Medicaid Services. Medicare Dialysis Facilities. 2022. https://data.cms.gov/quality-of-care/medicare-dialysis-facilities
26. Wolfe R, Ashby V, Milford E, et al. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *N Engl J Med*. 1999;341:1725-1730. doi:10.1056/NEJM199912023412303
27. United States Renal Data System. *2021 USRDS Annual Data Report: Epidemiology of Kidney Disease in the United States*. Bethesda, MD: National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases; 2021.
28. Marrero W, Naik A, Friedewald J, et al. Predictors of deceased donor kidney discard in the United States. *Transplantation*. 2017;101(7):1690-1697. doi:10.1097/TP.0000000000001238
29. Scientific Registry of Transplant Recipients. Technical methods for the program-specific reports. 2022. https://www.srtr.org/about-the-data/technical-methods-for-the-program-specific-reports/

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hartman N, He K. Individualized empirical null estimation for exact tests of healthcare quality. *Statistics in Medicine*. 2024;43(12):2403-2420. doi: 10.1002/sim.10074