

UC Irvine

UC Irvine Previously Published Works

Title

Impact of case-mix measurement error on estimation and inference in profiling of health care providers

Permalink

<https://escholarship.org/uc/item/9t2528cj>

Journal

Communications in Statistics - Simulation and Computation, 49(8)

ISSN

0361-0918

Authors

Şentürk, Damla
Chen, Yanjun
Estes, Jason P
[et al.](#)

Publication Date

2020-08-02

DOI

10.1080/03610918.2018.1515360

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed



Published in final edited form as:

Commun Stat Simul Comput. 2020 ; 49(8): 2206–2224. doi:10.1080/03610918.2018.1515360.

Impact of Case-Mix Measurement Error on Estimation and Inference in Profiling of Health Care Providers

Damla entürk¹, Yanjun Chen², Jason P. Estes³, Luis F. Campos⁴, Connie M. Rhee^{5,6}, Kamyar Kalantar-Zadeh^{5,6}, Danh V. Nguyen^{5,*}

¹Department of Biostatistics, University of California, Los Angeles, CA 90095, U.S.A.

²Institute for Clinical and Translational Science, University of California, Irvine, CA 92687, U.S.A.

³Research, Pratt & Whitney, East Hartford, CT 06118, U.S.A.

⁴Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

⁵Harold Simmons Center for Chronic Disease Research and Epidemiology, University of California Irvine School of Medicine, Orange, CA 92868, U.S.A.

⁶Department of Medicine, University of California Irvine, Orange, CA 92868, U.S.A.

Abstract

Profiling analysis aims to evaluate health care providers by modeling each provider's performance with respect to a patient outcome, such as unplanned hospital readmission. High-dimensional regression models are used in profiling to risk-adjust for patient case-mix covariates. Case-mix covariates typically ascertained from administrative databases are inherently error-prone. We examine the impact of case-mix measurement error (ME) on profiling models. The results show that even though the models' coefficient estimates are biased, this does not affect the estimation of standardized readmission ratio (SRR). However, ME leads to increased variation in SRR estimates and degrades the ability to identify under-performing providers.

Keywords

fixed effects; random effects; measurement error; hierarchical logistic regression; profiling analysis

1 Introduction

Profiling or evaluation of health care providers, including hospitals, nursing homes, and dialysis facilities, with respect to a patient outcome such as 30-day unplanned hospital readmission is important to ensure adequate and safe health care delivery. Profiling dates back nearly a century (Codman 1916) and has evolved over time to serve several purposes. This includes 1) identifying providers with below standard performance by government agencies for regulatory or payment purposes, 2) conveying information to patients regarding the quality of care, and 3) providing feedback to providers for quality improvement. In the

* danhvn1@uci.edu.

United States, a more systematic reporting of patient outcomes among providers has only appeared directly to consumers in the last decade by the Centers for Medicare and Medicaid Services (CMS). This includes condition-specific 30-day mortality (e.g., acute myocardial infarction, heart failure, pneumonia) and 30-day (all-cause) unplanned readmission rates; see Keenan et al. (2008), Krumholz et al. (2011), Lindenauer et al. (2011), and Horwitz et al. (2011, 2014).

In profiling analysis, there are two main goals: (a) report accurate estimates of provider effects relative to a reference, such as a national rate, and (b) identify providers that are “worse,” “better” or “not different” relative to the reference. We refer to goals (a) and (b) as estimation and inference goals, respectively. With respect to goal (a), each provider’s effect is summarized as a ratio in the form of a risk-standardized readmission ratio (SRR) for the patient outcome of 30-day unplanned readmission, for instance. Patient outcomes vary across providers due to variation in providers’ quality of care (provider effects) and variation in patient case-mix (patient-level effects, including demographics, comorbidities, and types of index hospitalization). With patients nested within providers, profiling models are hierarchical logistic regressions of the form *outcome* = *provider effects* + *patient case-mix effects*.

Because profiling models involve large populations of patients, such as all dialysis patients treated at dialysis facilities in the U.S. or all patients admitted to hospitals, the patient case-mix covariates are ascertained from administrative claims data. As such, measurement errors (MEs) in the case-mix covariates, such as inaccuracies in continuous variables and misclassified categorical variables (e.g., patient comorbidities) are expected and well-known. The impact of ME in the covariates is biased regression coefficient estimates, a problem that has been well-studied (Carroll et al. 2006). However, no methodological study has been conducted to understand the general effect of ME on the estimation of SRR and on the inferential objective of profiling. To date, one study has reported an empirical comparison of the random effects (RE)/CMS profiling model for 30-day readmission among patients with acute myocardial infarction, using variables based on claims data (imprecise) versus patient medical records (more precise). The study concluded that administrative claims data is suitable for profiling (Krumholz et al. 2011). Without knowledge of the statistical properties of profiling models under case-mix ME, the generalizability of conclusions from an empirical study is unclear. For instance, are the estimates of provider-specific SRRs biased and therefore invalid for public reporting when case-mix variables are measured with error? And, to what extent do case-mix MEs reduce profiling models’ ability to identify underperforming providers? We aim to fill this knowledge gap and provide more general insights on the impact of case-mix ME in profiling.

To investigate the impact of ME, our work here focuses on the outcome of 30-day unplanned hospital readmission; however, the methodological issues are general and applicable to other outcomes, such as all-cause or cause-specific in-hospital mortality. Also, we will consider both random effects (RE) and fixed effects (FE) profiling models proposed in the literature. We note that RE models have been the dominant approach and are the adopted method by CMS (CMS 2014; Ash et al. 2012; Krumholz et al. 2011). Motivations for hierarchical RE have largely been conceptual and center around the need to account for the nested structure

of the data (e.g., patients nested within providers) and data sparsity (low frequency of the observed outcome); see Normand and Shahian (2007), Normand et al. (1997), and Jones and Spiegelhalter (2011) for conceptual motivations for the RE models. (Throughout this work, we refer to the terms “RE model” and “CMS model” interchangeably, and they refer to the CMS adopted model with random intercepts for providers as further described in Section 2.)

FE models for profiling was proposed by Kalbfleisch and Wolfe (2013) and He et al. (2013), which model provider effects as fixed effects. We note that the FE model of He et al. (2013) is a *high-dimensional parameter model* with a unique fixed intercept for each provider. In practice, the number of providers is in the hundreds or thousands. The advantages of RE versus FE models have been discussed and compared previously (Kalbfleisch and Wolfe 2013; Chen et al. 2017); thus, it is not the focus of this paper. However, we briefly note that although RE models can provide stable provider effect estimates (through shrinkage), they are biased toward the overall provider average and the bias is larger for smaller providers (Kalbfleisch and Wolfe 2013). RE models also have smaller *overall average* estimation error, but this gain is in the center of the distribution of the outcomes, reducing the power to identify underperforming providers. We consider the impact of ME on both RE and FE models. Finally, to avoid confusion, we note that there is another use of the term “fixed effects” models in the profiling literature, where a “naive” FE model refers to a logistic regression model with one overall intercept (e.g., see Austin, Alter and Tu 2003). This naive FE model is no longer in wide usage and is different from the FE model of He et al. (2013).

2 Methods

2.1 RE and FE models without measurement error

The RE model implemented by CMS for (30-day unplanned) all-cause or condition-specific hospital readmission is the following RE logistic regression model,

$$g(\mu_{ij}^*) = \gamma_i^* + \beta^{*\top} \tilde{\mathbf{Z}}_{ij}, \quad \gamma_i \sim N(\gamma_0^*, \sigma^{*2}), \quad (1)$$

where $\mu_{ij}^* = p_{ij}^* = \Pr(Y_{ij} = 1 | \beta^*, \gamma_i^*, \tilde{\mathbf{Z}}_{ij})$ is the expected readmission for patient index discharge $j = 1, 2, \dots, n_i$ in provider $i = 1, 2, \dots, F$, and $Y_{ij} = 1$ if the j th index discharge at provider i results in a 30-day readmission, and equals 0 otherwise. Also, $g(p_{ij}^*) = \log\{p_{ij}^*/(1 - p_{ij}^*)\}$ is the logit function. Model (1) adjusts for patient case-mix variables, such as patient baseline and admission characteristics, denoted by the vector of r covariates $\tilde{\mathbf{Z}}_{ij} = (V_{ijr_1}, \dots, V_{ijr_1}, X_{ij1}, \dots, X_{ijr_2}) \equiv (\mathbf{V}_{ij}, \mathbf{X}_{ij})$ and their effects denoted by $\beta^{*\top} = (\beta_{v,1}^*, \dots, \beta_{v,r_1}^*, \beta_{x,1}^*, \dots, \beta_{x,r_2}^*)^\top \equiv (\beta_v^*, \beta_x^*)^\top$ (with $r = r_1 + r_2$).

In contrast, in the context of providers as dialysis facilities, Kalbfleisch and Wolfe (2013) and He et al. (2013) proposed modeling providers’ effects with fixed effects $(\gamma_1, \dots, \gamma_F)$:

$$g(\mu_{ij}^*) = \gamma_i^* + \beta^{*\top} \tilde{\mathbf{Z}}_{ij}, \quad i = 1, \dots, F. \quad (2)$$

For clarity, we emphasize that the FE model (2) is a single simultaneous model for all F providers with a high-dimensional parameter space and not a separate logistic regression model for each provider. We also note that for the FE model (2), in the context of performance assessment of dialysis facilities, further adjustment through inclusion of a hospital RE was considered in He et al. (2013), although it was found that the contribution of the hospital effect was small. In order to compare RE and FE models and for the results to be more applicable to different provider settings, we consider the FE model in (2) in this work.

2.2 Measurement error in patient case-mix variables

In practice, to date, it is assumed that the patient case-mix variables, $\tilde{\mathbf{Z}}_{ij}$, are measured without error and no methodological study has been conducted to examine the effect of ME on estimation and inference in profiling.

To examine the impact of ME on RE/FE profiling models (1) and (2), define the vector of observed case-mix variables $\mathbf{Z}_{ij} = (V_{ij1}, \dots, V_{ijr_1}, W_{ij1}, \dots, W_{ijr_2}) \equiv (\mathbf{V}_{ij}, \mathbf{W}_{ij})$ and their corresponding coefficients by $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_{v,1}, \dots, \boldsymbol{\beta}_{v,r_1}, \boldsymbol{\beta}_{w,1}, \dots, \boldsymbol{\beta}_{w,r_2})^T \equiv (\boldsymbol{\beta}_v, \boldsymbol{\beta}_w)^T$. For generality, let r_1 of the case-mix variables, namely $V_{ij\ell}$ $\ell = 1, \dots, r_1$, to be measured without ME, while a subset of r_2 case-mix variables ($W_{ij\ell}$ $\ell = 1, \dots, r_2$) are assumed to be ascertained with ME (e.g., patient comorbidities from administrative claims data). Thus, the observable RE and FE profiling models in practice are, respectively:

$$g(\mu_{ij}) = \gamma_i + \boldsymbol{\beta}^T \mathbf{Z}_{ij}, \quad \gamma_i \sim N(\gamma_0, \sigma^2), \quad (3)$$

and

$$g(\mu_{ij}) = \gamma_i + \boldsymbol{\beta}^T \mathbf{Z}_{ij}, \quad i = 1, \dots, F, \quad (4)$$

where μ_{ij} , γ_i , γ_0 , and σ^2 are analogously defined as above in Section 2.1.

Throughout we consider the classical ME model, $W_{ij\ell} = X_{ij\ell} + U_{ij\ell}$ where $U_{ij\ell}$'s are independent measurement errors in the ℓ th case-mix variable that is continuous. The degree of ME can be quantified by the reliability ratio $\lambda = \text{var}(X)/\text{var}(W) = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$. In the case of error-prone categorical case-mix variables, the ME is in the form of misclassification. For example, with the common binary patient comorbidities, consider the probability mass function of (X, W) : $p_{ab} \equiv \Pr(W_{ij\ell} = a | X_{ij\ell} = b)$ with $a, b = 0, 1$. The amount of ME for misclassification is summarized by the sensitivity and specificity of the case-mix variable $W_{ij\ell}$ respectively: $sn_{\ell} = \Pr(W_{ij\ell} = 1 | X_{ij\ell} = 1)$ and $sp_{\ell} = \Pr(W_{ij\ell} = 0 | X_{ij\ell} = 0)$.

It is well studied in the classical ME literature that for generalized linear models ME leads to biased estimates of coefficients (estimates of $\{\gamma_i, \boldsymbol{\beta}\}$ do not target $\{\gamma_i^*, \boldsymbol{\beta}^*\}$). However, how the quantities of interest in profiling, namely SRR, are impacted by ME is unknown. Also, how inference (e.g., identifying under-performing providers) is impacted by ME is also not known. We summarize these estimation and inference procedures for RE and FE models below.

2.3 Standardized readmission ratio

The main estimate of interest reported in profiling is the provider-specific standardized readmission ratio (SRR). The SRR for provider i is estimated as

$$SRR_i = \frac{\sum_{j=1}^{n_i} \hat{p}_{ij}}{\sum_{j=1}^{n_i} \hat{p}_{M,ij}}, \quad (5)$$

where $\hat{\gamma}_i$ is the provider effect estimate, $\hat{\beta}$ is the case-mix effect estimate, $\hat{p}_{ij} = g^{-1}(\hat{\gamma}_i + \hat{\beta}^T \mathbf{Z}_{ij})$ is the estimated probability of readmission for patient j in provider i and $\hat{p}_{M,ij} = g^{-1}(\hat{\gamma}_M + \hat{\beta}^T \mathbf{Z}_{ij})$. For the FE model, $\hat{\gamma}_M$ in the denominator is taken to be the median of the $\{\hat{\gamma}_i\}_{i=1}^F$ and for the RE model it is the estimated mean of the distribution of γ_i (namely $\hat{\gamma}_0$). The numerator of SRR_i is the expected total number of readmissions for facility i and the denominator is the expected total number of readmissions for an “average” provider (taken over the population of all providers), adjusted for the particular case-mix of the *same* patients in facility i . A SRR_i that is significantly larger or smaller than 1 indicates that provider i is under- or overperforming relative to the reference norm.

When there is no ME in the patient case-mix variables, the SRR is

$$SRR_i^* = \frac{\sum_{j=1}^{n_i} \hat{p}_{ij}^*}{\sum_{j=1}^{n_i} \hat{p}_{M,ij}^*}, \quad (6)$$

Where $\hat{p}_{ij}^* = g^{-1}(\hat{\gamma}_i^* + \hat{\beta}^{*T} \tilde{\mathbf{Z}}_{ij})$ and $\hat{p}_{M,ij}^* = g^{-1}(\hat{\gamma}_M^* + \hat{\beta}^{*T} \tilde{\mathbf{Z}}_{ij})$. In this case, SRR_i^* estimates the true/theoretical quantity $\overline{SRR}_i^* = \sum_{j=1}^{n_i} p_{ij}^* / \sum_{j=1}^{n_i} p_{M,ij}^*$, where $p_{ij}^* = g^{-1}(\gamma_i^* + \beta^{*T} \tilde{\mathbf{Z}}_{ij})$ and $p_{M,ij}^* = g^{-1}(\gamma_M^* + \beta^{*T} \tilde{\mathbf{Z}}_{ij})$, particularly for the FE model. (RE models provide biased estimates of SRR even in the absence of case-mix ME.) We illustrate in Section 4 that even though all model parameters (γ_M , $\{\gamma_i\}_{i=1}^F$, and β) are estimated with bias, this bias does not affect the estimation of SRR in either model.

2.4 Estimation and inference procedures

The RE model is a standard generalized linear mixed effects model for which available software can be used, including SAS PROC GLIMMIX or R library lme4 function glmer. The CMS implementation uses SAS PROC GLIMMIX (Ross et al. 2010; Horwitz et al. 2011). For the FE model estimation, He et al. (2013) proposed an iterative algorithm that alternates between estimation of $\{\gamma_i\}$ given β and estimation of β given $\{\gamma_i\}$ using one-step Newton-Raphson updates. Iteration terminates when $\max \|p_{ij}^{(t+1)} - p_{ij}^{(t)}\| < 10^{-6}$ on successive steps t . See He et al. (2013) for details. R codes for fitting both RE and FE models are provided as supplemental materials at <http://faculty.sites.uci.edu/nguyenlab/supplement/>. We note that in this work, references to ‘RE’ and ‘FE’ methods refer

specifically to the models described by (1 and 3) and (2 and 4) *together with the corresponding inference procedures* described below.

To understand the impact of ME on the ability to identify under-performing providers and providers not different from the national reference, we provide some details on the inference procedures for the RE and FE models. First, for the RE model, a bootstrap resampling of providers with replacement (500 samples) is used to obtain a 95% confidence interval (CI) for each \overline{SRR}_i . The REs sampled from the posterior distribution of γ_i are used to estimate \overline{SRR}_i in each bootstrap sample (Horwitz et al. 2011). Provider i is flagged as performing worse than expected, relative to the reference, if the lower confidence limit is above 1. Similarly, a provider is identified as performing better than the reference when the upper confidence limit is below 1. Providers with CI containing 1 are considered not different from the reference.

Steps of the RE inference procedure (Ash et al. 2012) are provided below. Note that the sampled provider random effect, $\hat{\gamma}_i^{(b)*}$ for facility i , is sampled from the posterior distribution in step 3 in each bootstrap iteration, which is based on case-mix variables with ME. As illustrated in Section 4, this results in wider confidence intervals on average.

RE Inference: Provider-Specific Bootstrap Confidence Interval

0. Fit the RE model (3). The provider-specific estimates are denoted by $\hat{\gamma}_i, i = 1, 2, \dots, F$, with overall mean $\hat{\gamma}_0$. Also, denote the variance and patient case-mix estimates by $\hat{\sigma}^2$ and $\hat{\beta}$, respectively. Calculate SRR_i as given by (5).
1. Generate a bootstrap dataset by sampling F providers with replacement from the original dataset. Denote the unique set of providers sampled by $\mathbb{F}^{(b)}$, where b indexes the bootstrap dataset.
2. Fit the RE model (3) to the bootstrap dataset and treat each resampled provider as distinct. Calculate:
 - (a) The patient case-mix effects, $\hat{\beta}^{(b)}$.
 - (b) The mean and variance of the distribution of the provider effects, $\hat{\gamma}_0^{(b)}$, and $\hat{\sigma}^{2(b)}$.
 - (c) The provider-specific effects and variances, $\{\hat{\gamma}_i^{(b)}, \widehat{\text{Var}}^{(b)}(\gamma_i)\}, i = 1, 2, \dots, F$. (If a provider is sampled more than once, then randomly select one set of the provider-specific estimates and variances.)
3. Generate a provider RE from the provider-specific distribution from step 2(c) for each unique provider, $i \in \mathbb{F}^{(b)}$. The posterior distribution of each RE is approximated by a normal distribution, $\hat{\gamma}_i^{(b)*} \sim N(\hat{\gamma}_i^{(b)}, \widehat{\text{Var}}^{(b)}(\gamma_i))$.
4. Calculate SRR_i for each unique provider i sampled in step 1:

RE Inference: Provider-Specific Bootstrap Confidence Interval

$$SRR_i^{(b)} = \sum_{j=1}^{n_i} \hat{p}_{ij}^{(b)} / \sum_{j=1}^{n_i} \hat{p}_{M,ij}^{(b)} = \sum_{j=1}^{n_i} g^{-1} \left(\hat{\gamma}_i^{(b)*} + \hat{\beta}^{(b)T} \mathbf{Z}_{ij} \right) / \sum_{j=1}^{n_i} g^{-1} \left(\hat{\gamma}_0^{(b)} + \hat{\beta}^{(b)T} \mathbf{Z}_{ij} \right), \text{ for } i \in \mathbb{F}^{(b)}.$$

5. Repeat the bootstrap procedure, step 1 - step 4, 500 times ($b = 1, \dots, 500$) and form the 95% CI estimate of \overline{SRR}_i for each provider $i = 1, 2, \dots, F$.
-

For the FE model, inference for \overline{SRR}_i can be equivalently based on the (fixed effect) provider specific parameters γ_i (for each $i = 1, \dots, F$). More specifically, FE inference is based on testing the hypothesis $H_0: \gamma_i = \gamma_M$ (i.e., $\overline{SRR}_i = 1$), accounting for the high-dimensional FE parameters. He et al. (2013) proposed a method based on resampling under the null hypothesis to evaluate the p-value, the “probability that a given provider would experience a number of readmissions as least as extreme as that observed if the null hypothesis is true, accounting for the provider’s patient case-mix.” Details are provided in the FE inference procedure below (He et al. 2013). In step 2, note that the Bernoulli draws $Y_{ij}^{(b)}$ under the null hypothesis are based on the median provider and patient case-mix effect estimates, $\hat{\gamma}_M$ and $\hat{\beta}$, estimated from case-mix variables with ME. This results in higher variance in the resampled test statistic, $Y_i^{(b)}$, under the null. As illustrated in Section 4, this results in reduced sensitivity to detect under-performing providers.

FE Inference: Provider-Specific Hypothesis Testing

1. Estimate FE model (4) and obtain estimates $\hat{\beta}$ and $\hat{\gamma}_M$.
 2. For the i th provider, draw $B = 500$ samples, $\{Y_{ij}^{(b)} : j = 1, 2, \dots, n_i\}_{b=1}^B$ under the null hypothesis, where each observation is independently drawn from a Bernoulli distribution: $Y_{ij}^{(b)} \sim \text{Ber}(\bar{p}_{ij})$, where $\bar{p}_{ij} = \exp(\hat{\gamma}_M + \hat{\beta}^T \mathbf{Z}_{ij}) / \{1 + \exp(\hat{\gamma}_M + \hat{\beta}^T \mathbf{Z}_{ij})\}$.
 3. Calculate the total number of readmissions in the resampled data: $Y_i^{(b)} = \sum_{j=1}^{n_i} Y_{ij}^{(b)}$.
 4. Calculate the p-value for testing $H_0 : \gamma_i = \gamma_M$ as follows. Compute $SL_i^+ \equiv B^{-1} \sum_{b=1}^B \left[0.5 \mathbb{1}(Y_i^{(b)} = O_i) + \mathbb{1}(Y_i^{(b)} > O_i) \right]$, where O_i is the observed number of readmissions for provider i in the original data and $\mathbb{1}(\cdot)$ denotes the indicator function; similarly, compute $SL_i^- \equiv B^{-1} \sum_{b=1}^B \left[0.5 \mathbb{1}(Y_i^{(b)} = O_i) + \mathbb{1}(Y_i^{(b)} < O_i) \right]$.
The p-value is $P = 2 \times \min\{SL_i^+, SL_i^-\}$.
 5. Repeat steps 2 – 4 for each provider, $i = 1, 2, \dots, F$.
-

3. Simulation studies

We consider 4 simulation models to address two specific aims: Determine the impact of patient case-mix ME on (1) estimation of SRR and (2) inference procedures in both RE and FE models. The studies provide answers to the following questions: (1) Does the inclusion of error-prone case-mix variables bias the estimation of provider-specific SRRs and as a consequence invalidate their interpretation/use? (2) How does the inclusion of error-prone case-mix variables affect inference, specifically, the ability to identify under-performing providers? The 4 simulation models (SM1-SM4) described below includes a simple set-up with two continuous variables (SM1), a more general dependence structure with 15 variables (SM2), a simple misclassification variable (SM3), and a more general model with 30 case-mix covariates based on the USRDS data that includes both ME on continuous variables and discrete variables.

3.1 Simulation model 1: Two continuous case-mix variables

We begin with the following simple, but instructive, basic profiling model without ME:

$$g(\mu_{ij}^*) = \gamma_i^* + \gamma_M^* + \beta_{v,1}^* V_{ij1} + \beta_{x,1}^* X_{ij1}, \quad (\text{SM1}^*)$$

with $i = 1, \dots, F = 1,000$ providers and two continuous case-mix variables, V_{ij1} and X_{ij1} , measured without error. For simplicity, we take both covariates to be normally distributed with means $\mu_V = 0$, $\mu_X = 2$, and variances $\sigma_V^2 = \sigma_X^2 = 1$, and with correlation $\rho(V_{ij1}, X_{ij1}) = 0.3$. The variable V_{ij1} is assumed to be measured without error. The unobserved variable is X_{ij1} and the corresponding observed variable is $W_{ij1} = X_{ij1} + U_{ij1}$, where $U_{ij1} \sim N(0, \sigma_u^2)$ as described in Section 2.2. Thus, the observable profiling model is

$$g(\mu_{ij}) = \gamma_i + \gamma_M + \beta_{v,1}V_{ij1} + \beta_{w,1}W_{ij1}. \quad (\text{SM1})$$

We considered two levels of ME by setting $\sigma_u^2 = 0.75^2$ (lower ME: $100 \times (1 - \lambda)\% = 36\%$) and $\sigma_u^2 = 2^2$ (higher ME: 80%). We parametrized model (SM1*) to include γ_M^* so that the baseline rates of readmission (BRR) can be varied. BRR of 14.3%, 27.3%, and 41.7% are referred to as low, medium and high, corresponding to $\gamma_M^* = \log(1/6)$, $\log(3/8)$, and $\log(5/7)$, respectively.

Provider-specific effects, γ_i^* , were set as follows. Among the 1,000 providers, 2.5% were under-performers (worse: $\mathcal{J}_W = \{1, \dots, 25\}$), 2.5% were over-performers (better: $\mathcal{J}_B = \{26, \dots, 50\}$), and the remaining 95% of providers, with effects not different (ND: $\mathcal{J}_{ND} = \{51, \dots, 1000\}$) from the national reference, were generated from a $N(0, \sigma^2)$ distribution with $\sigma^2 = 0.2^2$. Providers that were truly different from the national reference were generated as $\gamma_i^* \sim \text{Uniform}(0.4, 1.5)$ for $i \in \mathcal{J}_W$ and $\gamma_i^* \sim -\text{Uniform}(0.4, 1.5)$ for $i \in \mathcal{J}_B$. (We note here that when all provider effects were generated from $N(0, \sigma^2)$ and true outlying providers were defined as the 2.5% in the tails of the distribution, the results were similar.)

Because the impact of ME on profiling will depend on the provider effect size (P-ES) and the patient case-mix effect size (CM-ES), we considered two P-ES and three CM-ES settings. The need for this is clear when one considers the extreme cases. For example, if providers have large effects on whether patients have unplanned 30-day readmission relative to the contribution from patient factors, then ME on the patient case-mix variables will have negligible effects on the outcome. On the other hand, if patient risk factors and comorbidities are the dominating effects on readmission (relative to a small P-ES), then ME can have a large effect on the outcome. Therefore, we considered (P-ES 1): $\gamma_i^* \sim \text{Uniform}(0.4, 1.5)$ for $i \in \mathcal{J}_W$ and $\gamma_i^* \sim -\text{Uniform}(0.4, 1.5)$ for $i \in \mathcal{J}_B$; (P-ES 2): $\gamma_i^* \sim \text{Uniform}(0.6, 1.5)$, for $i \in \mathcal{J}_W$ and $\gamma_i^* \sim -\text{Uniform}(0.6, 1.5)$ for $i \in \mathcal{J}_B$ where provider signals have been increased. For patient case-mix effect size (CM-ES) we considered the following three increasing CM-ES settings: $\beta_{v,1}^* = \beta_{x,1}^* = 0.5, 1, \text{ and } 2$ (CM-ES1, CM-ES2, and CM-ES3, respectively).

We also considered the case where worse and better performing provider effects were generated as fixed constants across Monte Carlo datasets so that each provider's average bias in estimation can be characterized (see Section 4.1). For this, under the P-ES1 setting, the under-performing provider effects were set to $\gamma_i^* \in \{0.40, 0.46, \dots, 1.5\}$, for $i \in \mathcal{J}_W$. For over-performing providers, $\gamma_i^* \in \{-1.50, -1.45, \dots, -0.40\}$, for $i \in \mathcal{J}_B$. Similarly, under the P-ES2 setting: $\gamma_i^* \in \{0.60, 0.64, \dots, 1.50\}$, for $i \in \mathcal{J}_W$ and $\gamma_i^* \in \{-1.50, -1.46, \dots, -0.60\}$, for $i \in \mathcal{J}_B$.

The generated data consisted of provider volume ranging from 42 to 210 patients on average. More specifically, the number of patients were generated from a truncated Poisson

distribution, where $n_i = \sum_{h=1}^{1000} m_{ih} 1\{m_{ih} \leq 7\}$, where $m_{ih} \sim \text{Poisson}(15)$. This process mimics the sparse data structure of patient discharges across U.S. dialysis facilities; see He et al. (2013) for details. From n_i , we defined small, medium and large sized providers by tertile (small: 42–103; medium: 104–126; large: 127–210 patients on average). Two hundred datasets, each with 1,000 providers, were generated for each simulation study scenario.

3.2 Simulation model 2: More general dependence structure

We extended the simulation model (SM1*) to include 10 error-free case-mix variables and 5 variables with ME, a generalization that allowed for investigation of a more general dependence structure among all case-mix variables as is typical in real data applications. More specifically,

$$g(\mu_{ij}^*) = \gamma_i^* + \gamma_M^* + \beta_{v,1}^* V_{ij1} + \cdots + \beta_{v,10}^* V_{ij10} + \beta_{x,1}^* X_{ij1} + \cdots + \beta_{x,5}^* X_{ij5}, \quad (\text{SM2}^*)$$

with $i = 1, \dots, F = 1,000$ providers. As with model (SM1*), the observed variables with error are $W_{ij\ell} = X_{ij\ell} + U_{ij\ell}$, where $U_{ij\ell} \sim N(0, \sigma_u^2)$, for $\ell = 1, \dots, 5$, leading to the observable model

$$g(\mu_{ij}) = \gamma_i + \gamma_M + \beta_{v,1} V_{ij1} + \cdots + \beta_{v,10} V_{ij10} + \beta_{w,1} W_{ij1} + \cdots + \beta_{w,5} W_{ij5}. \quad (\text{SM2})$$

The case-mix vector, $\tilde{\mathbf{Z}}_{ij} = (\tilde{Z}_{ij1}, \dots, \tilde{Z}_{ij15}) = (\mathbf{v}_{ij}, \mathbf{X}_{ij})$ in (SM2*), was generated from a multivariate normal distribution with means 0, variances 1, and correlation $\rho_{\ell\ell'} \equiv \rho(\tilde{Z}_{ij\ell}, \tilde{Z}_{ij\ell'})$, $1 \leq \ell, \ell' \leq 15$. Using the magnitude of correlations observed in USRDS data as a guide, we considered a more general dependence among variables that was generated in 3 blocks with different correlation structures: $0.01 \leq \rho_{\ell\ell'} \leq 0.05$ in block 1 for variables \tilde{Z}_1 to \tilde{Z}_5 ; $0.05 \leq \rho_{\ell\ell'} \leq 0.1$ in block 2 for variables \tilde{Z}_6 to \tilde{Z}_{10} ; and $0.1 \leq \rho_{\ell\ell'} \leq 0.25$ in block 3 for variables \tilde{Z}_{11} to \tilde{Z}_{15} . Variables across blocks were also correlated in the range of $0.01 \leq \rho_{\ell\ell'} \leq 0.25$. Provider effects were the same as in model (SM1*). Low, medium and high CM-ES were set as: (CM-ES1) $\beta_{x,\ell} = 0.5$, (CM-ES2) $\beta_{x,\ell} = 1$, and (CM-ES3) $\beta_{x,\ell} = 2$, for $\ell = 1, \dots, 5$. Effects of the 10 error-free covariates were $\beta_{v,1}^* = \cdots = \beta_{v,5}^* = 0.25$ and $\beta_{v,6}^* = \cdots = \beta_{v,10}^* = 0.5$.

3.3 Simulation model 3: Simple case-mix misclassification

Simulation model 3 considers a simple error-prone categorical case-mix variable, where the ME is in the form of misclassification. Model 3 is as defined above for (SM1*):

$$g(\mu_{ij}^*) = \gamma_i^* + \gamma_M^* + \beta_{v,1}^* V_{ij1} + \beta_{x,1}^* X_{ij1}, \quad (\text{SM3}^*)$$

and

$$g(\mu_{ij}) = \gamma_i + \gamma_M + \beta_{v,1} V_{ij1} + \beta_{w,1} W_{ij1}, \quad (\text{SM3})$$

where $(X_{ij\ell}, W_{ij\ell})$ are binary covariates with joint distribution given by $p_{ab} = \Pr(W_{ij\ell} = a | X_{ij\ell} = b)$ with $a, b = 0, 1$. For low ME, we set $p_{00} = 0.42$, $p_{01} = 0.18$, $p_{10} = 0.12$, and $p_{11} = 0.28$;

thus, the case-mix classification sensitivity (sn) and specificity (sp) are $sn = 0.7$ and $sp = 0.6$. For the higher ME, we set $p_{00} = 0.15$, $p_{01} = 0.25$, $p_{10} = 0.5$, and $p_{11} = 0.1$, resulting in lower $sn = 0.167$ and $sp = 0.375$. The covariate, V_{ijl} , measured without error is correlated with X_{ijl} and generated through $(V_{ijl}|X_{ijl} = 0) \sim N(0, 1)$ and $(V_{ijl}|X_{ijl} = 1) \sim N(1, 1.5^2)$.

3.4 Simulation model 4: USRDS data characteristics

We also considered a fourth simulation study (SM4*/SM4) more tailored to the assessment of dialysis facilities. Paralleling works for profiling all-cause readmission for hospitals (CMS 2014; Horwitz et al. 2011), assessment of dialysis facilities included the following 30 patient case-mix covariates: Variables 1–4: age, body mass index (BMI), length of index hospitalization (days), time on dialysis (years); variable 5: sex; variables 6–7: high risk index hospitalization, diabetes as the cause of ESRD; and variables 8–30: 23 past-year comorbidities, which includes amputation status; chronic obstructive pulmonary disease; cardiorespiratory failure/shock; coagulation defects and other specified hematological disorders; drug and alcohol disorders; end-stage liver disease; fibrosis of lung or other chronic lung disorders; hemiplegia, paraplegia, paralysis; hip fracture/dislocation; major organ transplants; metastatic cancer; other hematological disorders; other infectious disease and pneumonias; other major cancers; pancreatic disease; psychiatric comorbidity; respirator dependence; rheumatoid arthritis and inflammatory connective tissue disease; seizure disorders; septicemia/shock; severe cancer; severe infection; and ulcers (CMS 2014). Our focus on ME is on the 23 past-year comorbidities, although for simplicity of exposition we also include variables 6–7 as potentially error-prone; thus, $\mathbf{X}_{ij} = X_{ij1}, \dots, X_{ij25}$.

To generate the binary covariates, \mathbf{X}_{ij} , we consider underlying latent continuous variables $\mathbf{L}_{ij} = (L_{ij1}, \dots, L_{ij25}) \sim N_{25}(\boldsymbol{\mu}_L, \boldsymbol{\Sigma}_L)$, with $\boldsymbol{\Sigma}_L = (\sigma_{\ell\ell}^2)$ and $\boldsymbol{\mu}_L = (\mu_1, \dots, \mu_{25})$ chosen to be the observed covariance matrix and the means/prevalences for binary covariates based on USRDS data, respectively. The binary covariates $X_{ij\ell}$, $\ell = 1, \dots, 25$, were then generated through the process: $X_{ij\ell} = 1\{L_{ij\ell} < Z_{\ell}\sigma_{\ell\ell} + \mu_{\ell}\}$, where $\mathbf{1}\{A\}$ is the indicator function for event A , $z_{\ell} = \Phi^{-1}(\mu_{\ell})$ and $\Phi^{-1}(\cdot)$ is the standard normal inverse CDF. This process generates binary covariates (e.g., patient comorbidities) with prevalences equal to the corresponding observed prevalences in the USRDS data (namely $\boldsymbol{\mu}_L$). Next, the observed binary covariates with ME, $\mathbf{W}_{ij1}, \dots, \mathbf{W}_{ij25}$, were then generated based on the joint distributions of $(X_{ij\ell}, W_{ij\ell})$ for a range of misclassification sensitivities and specificities (su_{ℓ}, sp_{ℓ}), as described in Section 2.2. The range of sn_{ℓ}, sp_{ℓ} used was from 0.5 to 0.8.

The remaining 4 continuous variables and sex (V_{ij1}, \dots, V_{ij5}) were generated from a $N_4(\boldsymbol{\mu}_V, \boldsymbol{\Sigma}_V)$ and a $Ber(0.48)$ distribution, where $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ were the observed mean and covariance based on the USRDS data. The distribution of facility effect sizes, $\{\gamma_i^*\}$, was modeled as $\gamma_i^* \sim N(0, 0.2^2)$ for facilities $i \in \mathcal{J}_{ND}$ and outlying facilities were generated as $\gamma_i^* \sim Uniform(0.6, 2)$ for $i \in \mathcal{J}_W$ and $\gamma_i^* \sim Uniform(-2, -0.6)$ for $i \in \mathcal{J}_B$ under the P-ES2 setting. For the P-ES1 setting, the distributions were $\pm Uniform(0.4, 2)$. The patient case-mix effects, $\boldsymbol{\beta}^{T*} = (\beta_1^*, \dots, \beta_{30}^*)$, were set to be proportional to the estimates based on the USRDS data.

4. Results

4.1 Estimation of SRR

We present results based on the simple simulation case with two continuous case-mix variables (SM1* and SM1) in more details since the results were similar across simulation studies. First, the model coefficients from both RE and FE models are biased. This includes the biased estimated coefficients for case-mix variables measured with or without ME ($\hat{\beta}_x, \hat{\beta}_z$) and provider effect estimates ($\hat{\gamma}_M, \hat{\gamma}_i, i = 1, \dots, 50$; Figure 1). Results given are for the case with $\beta_x = \beta_z = 2$. As expected, the magnitude of the bias in the provider-specific estimates, $\hat{\gamma}_i$, increases with larger case-mix effect size (CM-ES1 to CM-ES3) and with higher level of ME (low [36%] vs. high [80%] ME) for a given CM-ES; see Figure 2 for the FE model. More generally, the magnitude of estimation bias depends on the provider effect size, case-mix effect size and the level of ME. For instance, when the CM-ES is small relative to the provider-specific effect size, the impact of ME is not detectable (e.g., see row 1 of Figure 2). (Results are similar for the RE model; see supplemental Figure S1.) We note that although high-dimensional FE models with covariates measured with error has not been examined in the ME literature to date, biased coefficients are expected similar to ME in generalized linear models.

However, biased model coefficients due to ME do not affect the estimation of SRR. Figure 3 shows the average SRR estimates for truly outlying providers (under- and over-performing providers under P-ES1) with true SRR ranging from 0.6 to 1.4. For the FE model (Figure 3), the SRR estimates under ME (dashed line) coincides with SRR estimates when using the true case-mix variables without ME (dotted line). Because FE models provide (asymptotically) unbiased estimates, these coincide with the true SRR (solid gray line). For the RE model (Figure S2), the results show that SRR estimates under RE models with (dashed) and without (dotted) ME also coincide. However, we note that these RE estimates do not target the true SRR, particularly in the tails of the distribution of SRR, because RE are generally biased shrinkage estimates (Kalbfleisch and Wolfe, 2013), regardless of whether case-mix variables are measured with or without error. The results hold with low (36%) or high (80%) amount of ME. (The pattern of results are similar for larger provider effect size, P-ES2; not shown.)

Thus, biased estimates of case-mix effects and provider-specific estimates do not translate to biased SRR estimates. However, we note that the variance in the estimates of SRR are higher under case-mix ME (results not shown). Increased variation as a consequence of covariate ME in regression models is also known in the classical ME setting (Carroll et al. 2006).

4.2 Inference: Identifying outlying providers

4.2.1 Overall impact of measurement error on profiling—To describe the impact of case-mix ME on profiling, we focus on the primary profiling goal of correctly identifying providers that under-perform (sensitivity “worse”: SEN-W) relative to the reference standard, and specificity (SPEC). SPEC refers to the correct identification of providers that are not different (ND) compared to the reference. Because provider assessment policies

focus on identifying under-performing providers, we focus on the results for SEN-W and SPEC-ND. (Sensitivity “better” was similar to SEN-W in our studies due to symmetry.)

Overall, the studies consistently found that although the estimation of SRR is not affected by ME, both FE and RE models have substantially reduced power to detect truly underperforming providers. Because the overall patterns of results are similar across studies, we present in more detail the results for model SM1 with 2 continuous case-mix covariates. Figure 4 illustrates the general patterns of the impact of ME on the both FE and RE models. Row 1 of Figure 4 shows the average SEN-W for identifying under-performing providers as a function of case-mix effect size ($\beta = 0.5, 1, 2$) for the case of low ME. The results support the following with respect to both FE and RE models: (1) The presence of case-mix ME does negatively impact profiling. (2) The reduction in the ability to identifying under-performing providers due to ME depends on the CM-ES. (3) These results are consistent across the two P-ES settings (low: black, high: gray curves) and also for the high level of ME (not shown). For example, average SEN-W was 74.3% and 62.0% for FE and RE models without ME, respectively. These SEN-W averages dropped to 69.2% and 52.3%, respectively, for FE and RE models with ME. Furthermore, note that when the case-mix effect is smallest (i.e., $\beta = 0.5$), average sensitivities were not different between the models using case-mix variables with and without ME. However, this should not be assumed to be the case in practice because case-mix variables, such as comorbidities, can have large effects on a patient’s likelihood of readmission.

The specificity (SPEC-ND), i.e., rate of correctly identifying the providers that do not differ from the reference standard, as a function of CM-ES is summarized in Figure S3 for the case of low ME. As expected, the reduction in sensitivity for identifying outlying providers in the presence of ME led to a slight increase in SPEC-ND for FE model (e.g., 93.2% vs. 93.9%: w/o vs. w/ME at P-ES1) and RE model (97.2% vs. 98.2%: w/o vs. w/ME at P-ES1).

We note that the overall average flagging performance/sensitivity patterns (with and without ME) as a function of case-mix effect size are similar when examined separately for providers with small, medium, and large volume. This is illustrated in Figure S4 for both FE and RE models under low provider effects size (P-ES1).

Finally, we also examined the confidence interval length and the test level (coverage probability [CP]) for the RE and FE inference procedures, respectively. For the FE model, Figure 5 (row 1) shows the CPs for the hypothesis testing procedure ($H_0: \gamma_i = \gamma_M$) under no ME and under ME with increasing CM-ES in both high and low levels of ME. CPs do not target the 0.95 level and are further below 0.95 for higher ME and larger CM-ES, as expected. For the RE model, the inference procedure based on 95% CI yielded wider CI length on average under case-mix ME (Figure 5, row 2).

4.2.2 More general dependence/correlation and case-mix misclassification—

To examine whether the adverse impact of ME holds under a more general profiling model, we consider a model containing 15 continuous case-mix variables with dependence/correlation within and across blocks of variables. The correlation structure consists of *unequal* correlations among patient case-mix variables (0.01 ρ_{ij} 0.25) selected to be

similar to USRDS data described in Section 3.1. As illustrated in Figure S5 - column 1, the patterns of results for both RE and FE models were similar to the 2 variables case described above; however, in this more realistic model, the difference in average SEN-W between the models with and without ME is much greater. For example, at CM-ES $\beta = 2$ under low ME, the SEN-W for FE models with vs. without ME were 36.0% vs. 55.3% (and similarly, for RE model with vs. without ME: 6.7% vs. 27.9%). The results were similar for high level of ME (FE model with vs. without ME: 27.3% vs. 56.3%; RE model with vs. without ME: 2.1% vs. 25.4%).

Also, Figure S5 - column 2, illustrates the impact of ME in categorical variables, i.e., misclassification of categorical case-mix variables. The impact of misclassification on profiling in this case is similar to the results described earlier for two continuous covariates (Figure 4).

4.2.3 Simulation model based on USRDS data—Profiling results for simulated data modeled after USRDS data are summarized in Figure 6 for baseline readmission rate (BRR) of 27.3% (medium). The results are similar for lower (14.3%) and higher (41.7%) BRR. The general patterns of the impact of ME on the ability to identify truly under-performing providers (SEN-W) is similar to the those described earlier, although the loss in power is more severe. For example, the overall average SEN-W was 85.9% for the FE model without ME and was reduced to an average of 64.1% for case-mix measured with error. For the RE model, the effect of ME was more severe with an average SEN-W of 79.1% vs. 37.8% for models with and without ME, respectively. Not surprisingly, the direction of SPEC-ND is reversed (with ME > without ME) because the majority ($190,000 = 200 \times 1000 \times 0.95$) of providers are truly ND relative to the reference standard.

5 Discussion

In this work we presented the first systematic study of the impact of measurement error on patient case-mix variables in profiling models. We found that estimates of SRR are valid under ME; therefore, their use is acceptable even under case-mix variables measured with error. However, ME increases variation in the SRR estimates and degrades the power of profiling methods to detect under-performing providers. Although beyond the scope of this work, careful studies to examine/document the degree of ME for all case-mix covariates (e.g., all comorbidities) is warranted to assess their impact. Furthermore, investigation of novel methods to improve the profiling performance under ME is needed. Finally, we note that our work focused on simulation models to understand the basic impact of measurement error on estimation and inference in profiling models. Further studies providing a more thorough statistical/theoretical treatment will contribute more general insights that may lead to additional guidance on the use of profiling models in the context of case-mix measurement error.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was supported by NIDDK grants R01 DK092232 and K23 DK102903. The interpretation/reporting of the data presented are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the U.S. government. We are grateful for comments from anonymous reviewers.

References

- Ash AS, Fienberg SE, Louis TA, Normand ST, Stukel TA, and Utts J 2012 Statistical issues in assessing hospital performance The COPSS-CMS White Paper Committee, CMS, Washington D.C.
- Austin PC, Alter DA, and Tu JV 2003 The use of fixed- and random-effects models for classifying hospitals as mortality outliers: a Monte Carlo assessment. *Medical Decision Making* 23:526–539. [PubMed: 14672113]
- Carroll RJ, Ruppert D, Stefanski LA and Crainiceanu CM 2006 Measurement error in nonlinear models: A modern perspective. Boca Raton: Chapman and Hall/CRC.
- Centers for Medicare & Medicaid Services (CMS)/UM-KECC. 2014 Report for the standardized readmission ratio. CMS, Washington, D.C.
- Chen Y, Senturk D, Estes JP, Campos LF, Rhee CM, Dalrymple LS, Zhang L, Kalantar-Zadeh K, Nguyen DV 2017 Performance characteristics of profiling methods and the impact of inadequate case-mix adjustment Submitted for publication.
- Codman E 1916 Hospitalization standardization. *Surgery, Gynecology, and Obstetrics* 22:119–120.
- He K, Kalbfleisch JD, Li Y, and Li Y 2013 Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. *Lifetime Data Analysis* 19:490–512. [PubMed: 23709309]
- Horwitz L, Partovain C, Lin ZQ, Herrin J, Grady J, Conover M, Montague J, Dillaway C, Bartczak K, Ross J, et al. 2011 Hospital-wide (all-condition) 30 day risk-standardized readmission measure CMS, Washington, D.C.
- Horwitz L, Partovain C, Lin ZQ, Grady J, Herrin J, Conover M, Montague J, Dillaway C, Bartczak K, Suter LG, et al. 2014 Development and use of an administrative claims measure for profiling hospital-wide performance on 30-day unplanned readmission. *Annals of Internal Medicine* 161:S66–75. [PubMed: 25402406]
- Jones HE, and Spiegelhalter DJ 2011 The identification of “unusual” health-care providers from a hierarchical model. *American Statistician* 65:154–163.
- Kalbfleisch JD, and Wolfe RA 2013 On monitoring outcomes of medical providers. *Statistics in Biosciences* 5:286–302.
- Keenan PS, Normand SL, Lin Z, Drye EE, Bhat KR, Ross JS, Schuur JD, Stauffer BD, Bernheim SM, Epstein AJ, et al. 2008 An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. *Circulation Cardiovascular Quality and Outcomes* 1:29–37. [PubMed: 20031785]
- Krumholz HM, Lin Z, Drye EE, Desai MM, Han HF, Rapp MT, Mattera JA, and Normand S-L 2011 An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. *Circulation Cardiovascular Quality and Outcomes* 4:243252.
- Lindenauer PK, Normand SL, Drye EE, Lin Z, Goodrich K, Desai MM, Bratzler DW, O’Donnell WJ, Metersky ML, Krumholz HM 2011 Development, validation, and results of a measure of 30-day readmission following hospitalization for pneumonia. *Journal of Hospital Medicine*, 6:142–150. [PubMed: 21387551]
- Normand ST, Glickman ME, and Gatsonis CA 1997 Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association* 92:803–814.
- Normand ST, and Shahian DM 2007 Statistical and clinical aspects of hospital outcomes profiling. *Statistical Science* 22:206–226.
- Ross JS, Normand SL, Wang Y, Ko DT, Drye EE, Keenan PS, Lichtman JH, Bueno H, Schreiner GC, and Krumholz HM 2010 Hospital volume and 30-day mortality for three common medical conditions. *New England Journal of Medicine* 362:1110–1118.

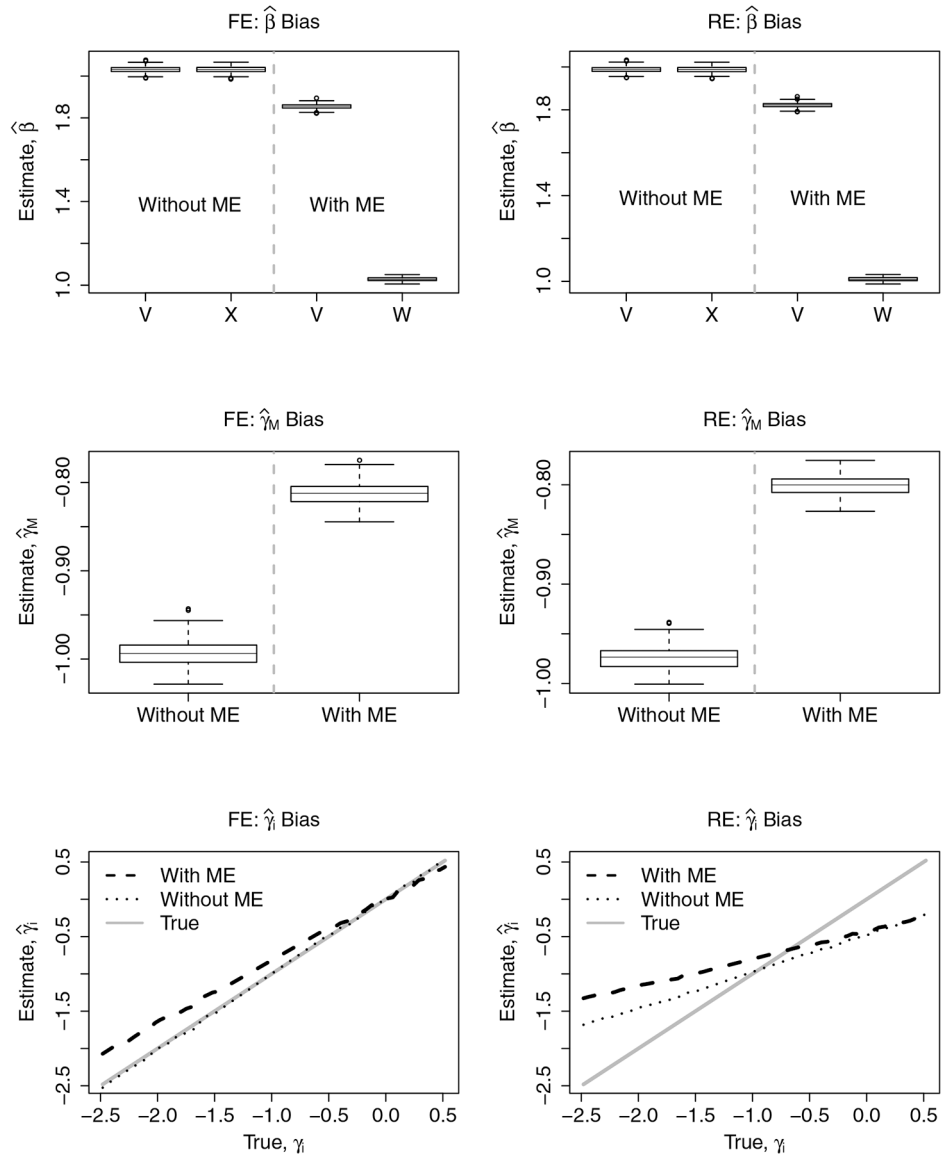


Figure 1: Bias case-mix coefficients estimates ($\hat{\beta}_z, \hat{\beta}_x$; row 1), overall “mean” provider effect estimates ($\hat{\gamma}_M$; row 2), and provider-specific estimates ($\hat{\gamma}_i$; row 3) due to measurement error for FE and RE models. (FE: fixed effects; RE: random effects; ME: measurement error)

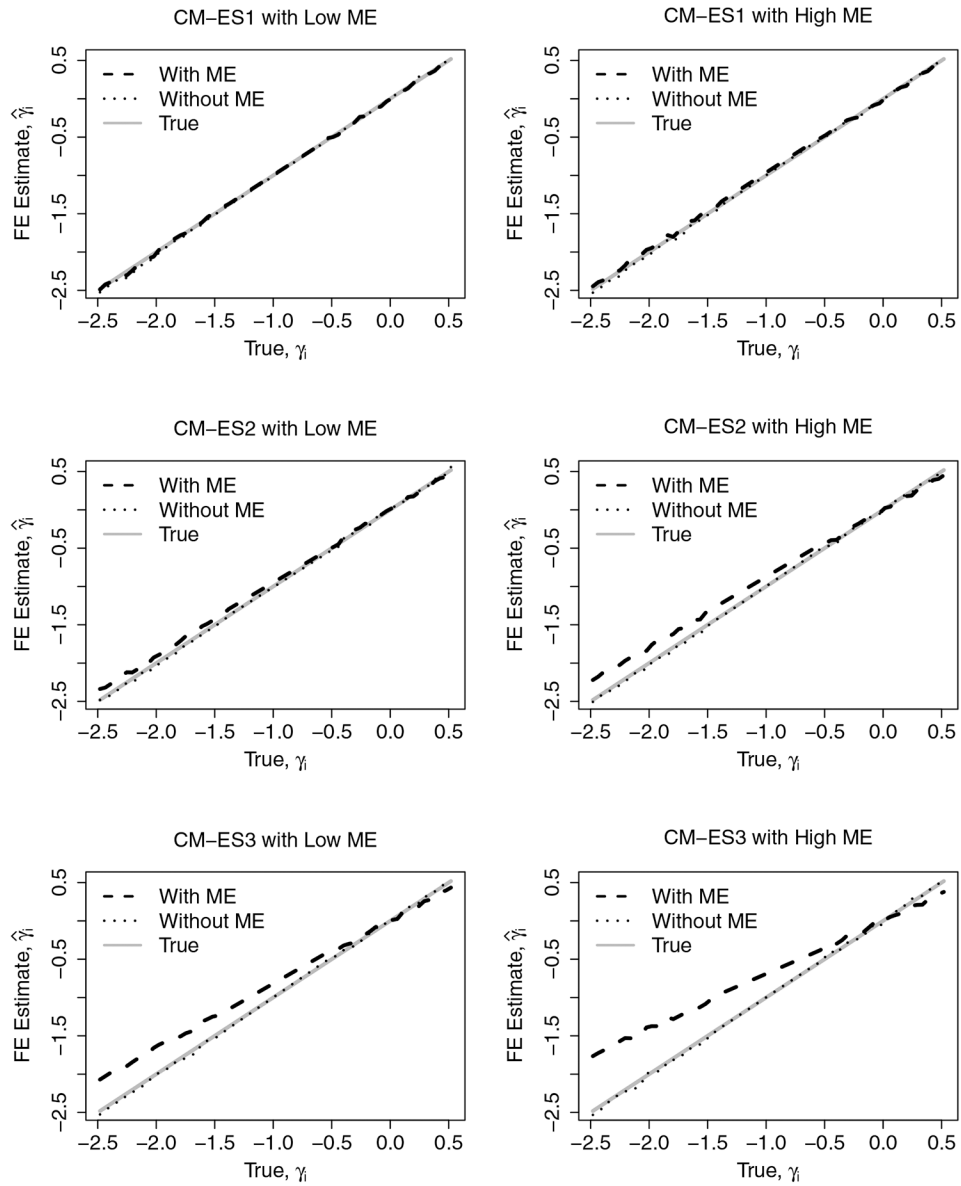


Figure 2: Bias of provider-specific estimates ($\hat{\gamma}_i$) of fixed effects (FE) models as a function of case-mix effect size (CM-ES 1 to CM-ES 3 corresponding to $\beta_z = \beta_x = 0.5, 1, \text{ and } 2$) and the level of measurement error (ME: low or high ME).

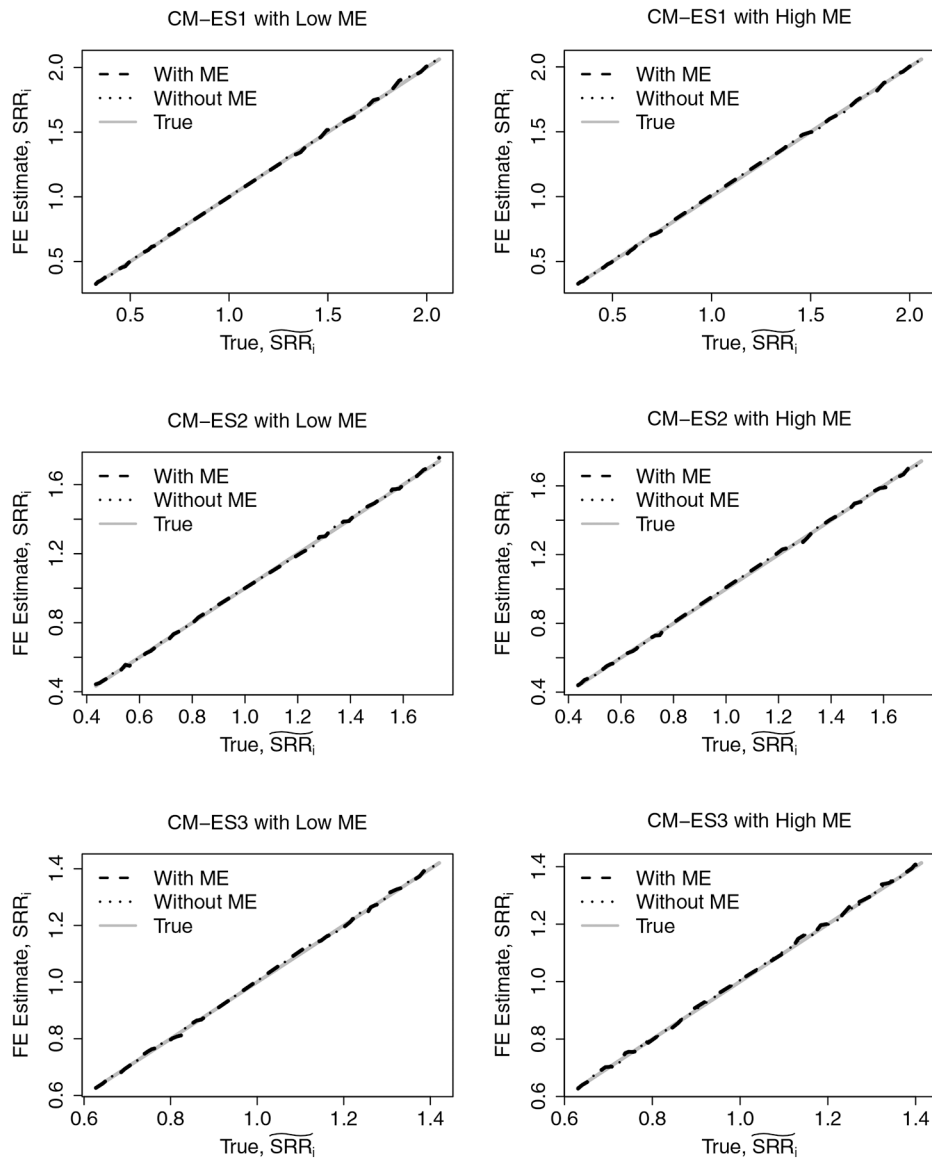


Figure 3: Provider-effect estimates, SRR_i , for fixed effects (FE) models with and without measurement error (column 1 - low ME, column - high ME) and across case-mix effect sizes (CM-ES).

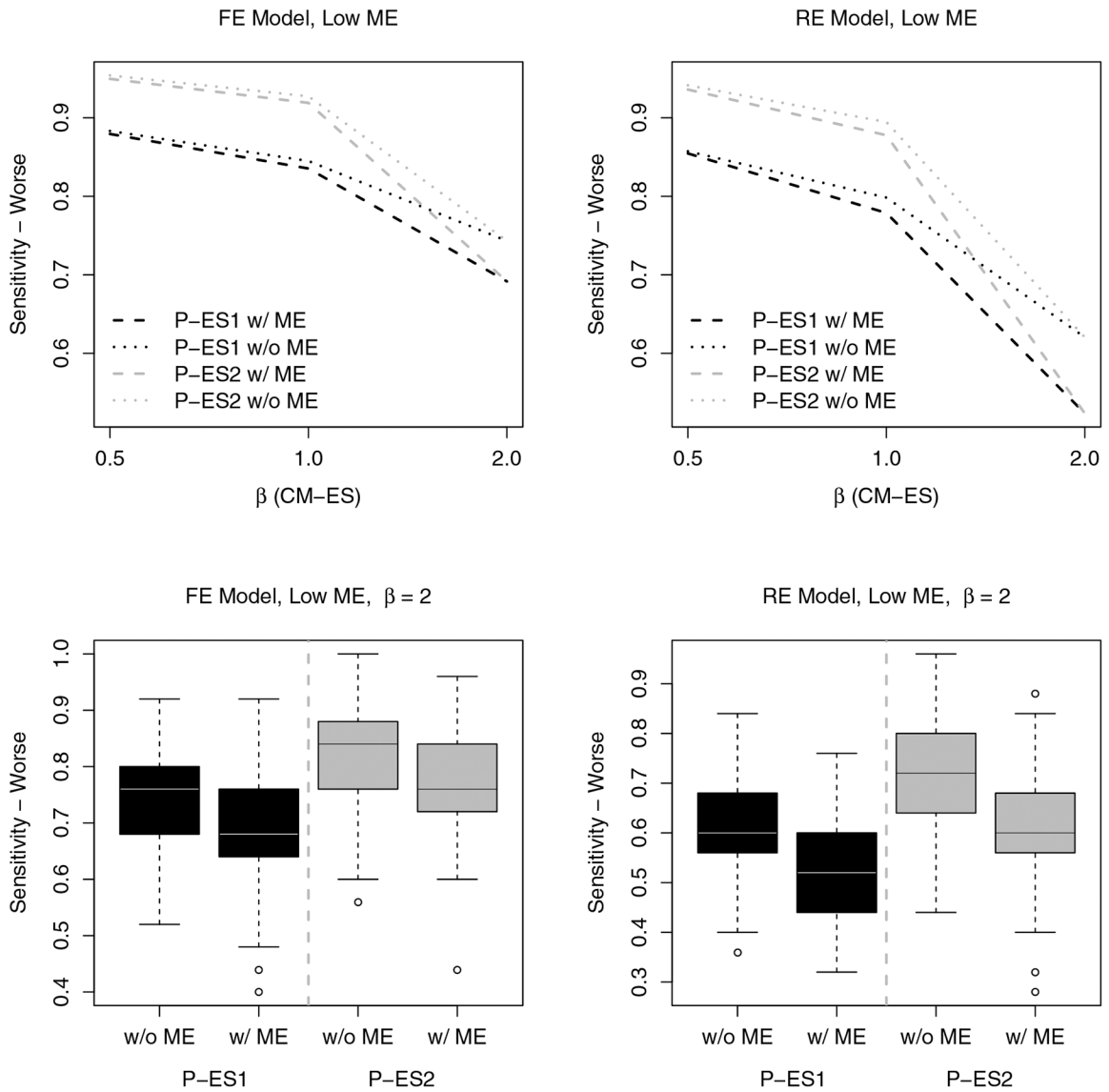


Figure 4: Row 1: Overall average flagging performance/sensitivity for identifying underperforming providers as a function of case-mix effect size (CM-ES: $\beta= 0.5,1, 2$), averaged over 200 simulated datasets. Given are results for FE and RE models when patient case-mix variables are ascertained without measure error (w/o ME; dotted), with ME (w/ME; dashed), and for provider effect sizes (P-ES1, P-ES2: low, high). Row 2: Distribution of sensitivity for identifying truly under-performing providers for the case of $\beta = 2$.

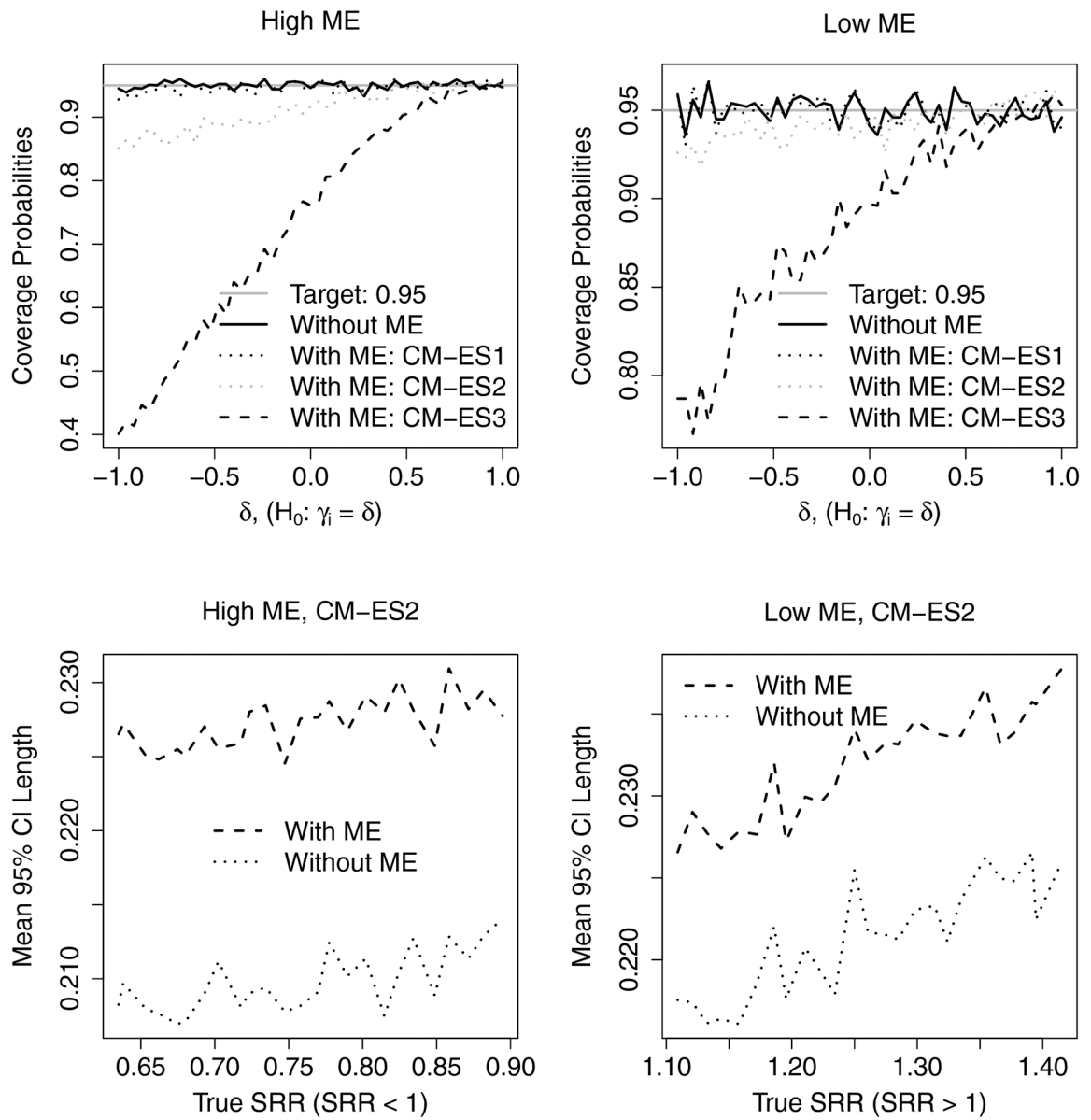


Figure 5: Coverage probabilities for hypothesis testing procedure use in fixed effects models (row 1) based on 1,000 Monte Carlo datasets. Average length of 95% confidence interval (CI) for SRR based on random effects inference procedure (row 2). (CM-ES: case-mix effect sizes 1–3 are $\beta=0.5, 1, 2$; ME: measurement error; SRR: standardized readmission ratio)

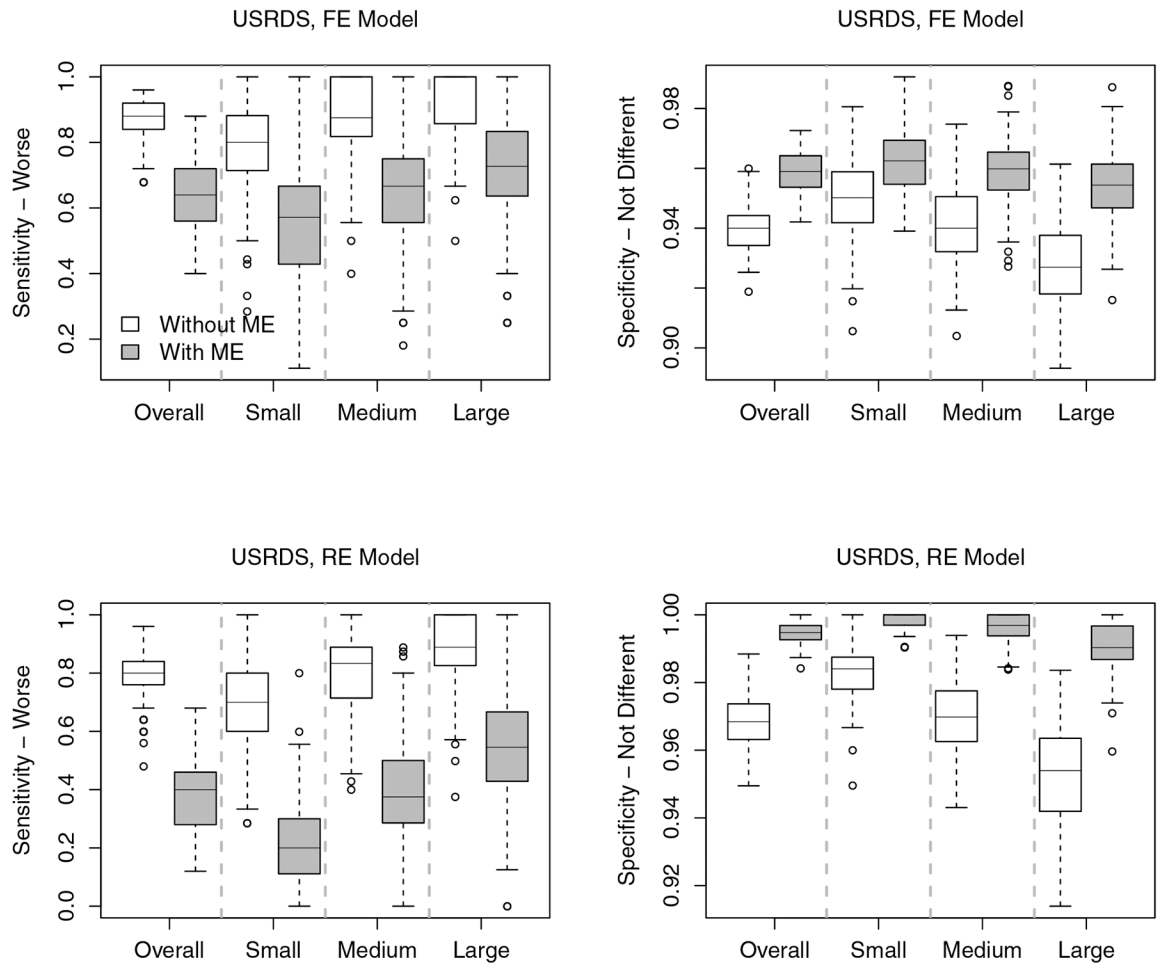


Figure 6: Performance of profiling models under USRDS data simulation studies: Fixed effects (FE) model (row 1) and random effects (RE) model (row 2).

FE Inference: Provider-Specific Hypothesis Testing

1. Estimate FE model (4) and obtain estimates $\hat{\beta}$ and $\hat{\gamma}_M$.
 2. For the i th provider, draw $B = 500$ samples, $\{Y_{ij}^{(b)} : j = 1, 2, \dots, n_i\}_{b=1}^B$ under the null hypothesis, where each observation is independently drawn from a Bernoulli distribution: $Y_{ij}^{(b)} \sim \text{Ber}(\bar{p}_{ij})$, where $\bar{p}_{ij} = \exp(\hat{\gamma}_M + \hat{\beta}^T \mathbf{Z}_{ij}) / \{1 + \exp(\hat{\gamma}_M + \hat{\beta}^T \mathbf{Z}_{ij})\}$.
 3. Calculate the total number of readmissions in the resampled data: $Y_i^{(b)} = \sum_{j=1}^{n_i} Y_{ij}^{(b)}$.
 4. Calculate the p-value for testing $H_0 : \gamma_i = \gamma_M$ as follows. Compute $SL_i^+ \equiv B^{-1} \sum_{b=1}^B [0.5 \mathbb{1}(Y_i^{(b)} = O_i) + \mathbb{1}(Y_i^{(b)} > O_i)]$, where O_i is the observed number of readmissions for provider i in the original data and $\mathbb{1}(\cdot)$ denotes the indicator function; similarly, compute $SL_i^- \equiv B^{-1} \sum_{b=1}^B [0.5 \mathbb{1}(Y_i^{(b)} = O_i) + \mathbb{1}(Y_i^{(b)} < O_i)]$.
The p-value is $P = 2 \times \min\{SL_i^+, SL_i^-\}$.
 5. Repeat steps 2 – 4 for each provider, $i = 1, 2, \dots, F$.
-