# Do Common Risk Adjustment Methods Do Their Job Well if Center Effects are Correlated With the Center-Specific Mean Values of Patient Characteristics?

Vach, Werner; Wehberg, Sonja; Luta, George

*Published in:*
Medical Care

*DOI:*
10.1097/MLR.0000000000002008

*Publication date:*
2024

*Document version:*
Final published version

*Document license:*
CC BY-NC-ND

*Citation for pulished version (APA):*
Vach, W., Wehberg, S., & Luta, G. (2024). Do Common Risk Adjustment Methods Do Their Job Well if Center Effects are Correlated With the Center-Specific Mean Values of Patient Characteristics? *Medical Care*, *62*(11), 773-781. https://doi.org/10.1097/MLR.0000000000002008

OPEN

# Do Common Risk Adjustment Methods Do Their Job Well If Center Effects Are Correlated With the Center-Specific Mean Values of Patient Characteristics?

*Werner Vach, PhD,*† Sonja Wehberg, PhD,‡ and George Luta, PhD§||¶*

**Background:** Direct and indirect standardization are well-established approaches to performing risk adjustment when comparing outcomes between healthcare providers. However, it is an open question whether they work well when there is an association between the center effects and the distributions of the patient characteristics in these centers.

**Objectives and Methods:** We try to shed further light on the impact of such an association. We construct an artificial case study with a single covariate, in which centers can be classified as performing above, on, or below average, and the center effects correlate with center-specific mean values of a patient characteristic, as a consequence of differential quality improvement. Based on this case study, direct standardization and indirect standardization—based on marginal as well as conditional models—are compared with respect to systematic differences between their results.

**Results:** Systematic differences between the methods were observed. All methods produced results that partially reflect differences in mean age across the centers. This may mask the classification as above, on, or below average. The differences could be explained by an inspection of the parameter estimates in the models fitted.

**Conclusions:** In case of correlations of center effects with center-specific mean values of a covariate, different risk adjustment methods can produce systematically differing results. This suggests the routine use of sensitivity analyses. Center effects in a conditional model need not reflect the position of a center above or below average, questioning its use in defining the truth. Further empirical investigations are necessary to judge the practical relevance of these findings.

**Key Words:** risk adjustment, provider comparisons, direct and indirect standardization, correlation, systematic differences

(*Med Care* 2024;62:773–781)

Quality comparisons between healthcare providers are of interest to different stakeholders. Patients may seek the best local provider, relevant authorities or insurance companies may want to identify poor-performing providers, provider networks may want to learn from differences in quality, and quality improvement programs may want to identify the best providers to use as role models. Consequently, comparing quality indicators between healthcare providers is part of quality monitoring or reimbursement strategies in many healthcare systems.[1–4] As patient populations often differ between providers with respect to the distribution of risk factors, a simple comparison of outcome frequencies between providers contradicts the ultimate goal of quality assurance in healthcare, that is, to accurately measure quality and thereby foster improvement.[5] Consequently, risk adjustment is mandatory.[6–8] Both direct and indirect standardization allow a risk-adjusted comparison across centers.[9–12] For the usual case of binary outcomes, both approaches are typically based on fitting a logistic regression model with potential risk factors as covariates. There has been an extensive discussion about the pros and cons of modeling center effects as fixed or random effects[13–17] and about how to incorporate the output of such models into indirect standardization.[18] However, in recent years, the question of robustness against violation of modeling assumptions has also attracted attention. Topics addressed were the impact of measurement error,[19,20] variation of regression coefficients across centers,[21] and the choice of the link function[22] on provider ranking and provider classification. There have also been attempts to develop methods that do not require a correct model specification but still allow a causal interpretation.[23,24]

One specific topic addressed in this context is the potential association between the center effects and char-

From the *Basel Academy for Quality and Research in Medicine, Basel, Switzerland; †Department of Environmental Sciences, University of Basel, Basel, Switzerland; ‡The Research Unit of General Practice, Department of Public Health, University of Southern Denmark, Odense, Denmark; §Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC; ||Clinical Research Unit, The Parker Institute, Copenhagen University Hospital, Bispebjerg and Frederiksberg, Nordre Fasanvej, Frederiksberg, Denmark; and ¶Department of Clinical Epidemiology, Aarhus University, Olof Palmes Allé, Aarhus, Denmark.

The authors declare no conflicts of interest.

Correspondence to: Werner Vach, PhD, Department of Environmental Sciences, University of Basel, Spalenring 145, CH-4055 Basel, Switzerland. E-mail: werner.vach@unibas.ch.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.lww-medicalcare.com.

DOI: 10.1097/MLR.0000000000002008

acteristics (especially the mean value) of the distributions of patient characteristics in these centers. Simulation-based investigations have been presented[15,25] or the association has been mentioned as a potential issue.[18] Such an association can arise, for example, when well-performing centers specialize in complex patients, or when centers serving more complex patients start investing more in quality improvement. In this article, we focus on the question of whether the presence of such an association can lead to systematic differences between results obtained from direct standardization and 2 common variants of indirect standardization.

It is not trivial to answer this question. If different risk adjustment methods give systematically different results, then we need to know some truth in order to identify the most adequate method. Traditionally, the truth is defined by the center effects in the data-generating logistic regression model. However, we will demonstrate that this approach may not be adequate in this context.

Consequently, we need an alternative framework to define some "truth" independent of such a model. We approach this by constructing an artificial case study in which expert or layperson audiences may intuitively agree on a (partial) ranking of the quality of providers without relying on a specific model.

Our interest is in systematic differences between the results of the risk adjustment. Such systematic differences should not depend on the sample size available in each center. Hence, the investigations will be based on assuming a very large sample size within each center. Furthermore, we are interested in systematic differences of non-negligible magnitude, in the sense that they have the potential to imply substantially different conclusions. They should go beyond numerical differences we have to expect simply because the methods use different formulas.

We start with a review of methods for direct and indirect standardization, focusing on a description of their intended role and a mathematically precise description of their formal approach. We then present an artificial case study illustrating how a correlation between center effects and the center-specific mean values of a patient characteristic can arise in a healthcare system. We apply the different risk adjustment methods to this case and then try to find explanations for the observed differences between the results. Finally, we suggest four types of actions implied by the findings of this article, including the routine check of sensitivity of results to the choice of the risk adjustment method.

## METHODS FOR DIRECT AND INDIRECT STANDARDIZATION

The general setting of risk adjustment for a provider comparison can be described by 3 random variables defined at the patient level in a relevant patient population, for example, all patients in a country with a specific diagnosis or receiving a specific treatment. These random variables are:

$Y$ A binary outcome variable indicates an unfavorable event.
$C$ The provider (center) taking care of the patient.
$X$ A vector of covariates reflecting potential risk factors for the outcome is present at baseline.

The need for risk adjustment arises if the distribution of $X$ varies across the centers. In this case, empirical frequencies of the unfavorable event, that is, empirical estimates of the raw prevalence values

$$\pi_c = P(Y = 1|C = C)$$

do not constitute a valid basis for performance comparisons across providers. They may partially reflect differences in patient populations.

Direct and indirect standardization aim at providing center-specific values which allow a meaningful comparison between providers in the case of differences between patient populations. A link to a reference population is used to reach this aim. Both approaches date back to the 19th century,[26] a long time before the invention of logistic regression. However, today logistic regression models constitute a basic ingredient to these approaches. We describe in the sequel these modern variants, typically used today for provider comparisons.

### Direct Standardization

Direct standardization estimates the relationship between patient characteristics and outcome in the center $c$ of interest, and then applies this relationship to all patients from a reference population to obtain estimates of the individual risk. The average over all these estimates can then be interpreted as the expected overall prevalence if all patients from the reference population would have been treated by center $c$. These values promise to allow a fair comparison between centers as they all refer to the same patient population. In this article, we will only consider the case of using all patients from all providers as the reference population.

Formally, this approach can be based on estimating

$$p_c(x) = P(Y = 1 \mid C = c, X = x).$$

If such an estimate $\hat{p}_c$ is available, direct standardization can be performed by approximating

$$d_c = E[p_c(X)]$$

via

$$\hat{d}_c = \frac{1}{N}\sum_{i}^{N}\hat{p}_c(X_i),$$

with $X_i$ denoting the covariate values of patient $i$ among the $N$ patients in the whole patient population.

Typically, estimation of $p_c(x)$ is based on a logistic regression model with center effects $\xi_c$. This model can be represented as

$$\operatorname{logit} p_c(x) = \xi_c + g_\beta(x),$$

with $g_\beta(x)$ denoting a prespecified function of the covariates and some regression parameters $\boldsymbol{\beta}$. A simple, typical choice is $g_\beta(x) = \beta_0 + \sum_j \beta_j x_j$. In the statistical literature, such a model is often called a *conditional* model (because of the conditioning on $C = c$) in contrast to a *marginal* model (cf. next subsection). For this reason, we denote the regression coefficients from this model with $\beta^C$. Moreover, for the conditional model, we assume in what follows that $g_\beta(x)$ includes an intercept and the values $\xi_c$ are centered around 0.

It had also been suggested to directly use the center effects $\xi_c$ to compare the centers.[6,18,27] This leads to the same ranking of centers as $d_c$, as the latter is a monotone transformation of $\xi_c$. However, $d_c$ may be easier to communicate, as it can be interpreted as a prevalence.

## Indirect Standardization

Indirect standardization estimates the relationship between patient characteristics and outcome in a reference population and applies this relationship to all patients from center $c$ to obtain estimates of the individual risk. The average over all these estimates can then be interpreted as the expected prevalence $e_c$ in center $c$, if the outcome of the patients from center $c$ follows the relationship observed in the reference population. By applying the same relationship to all centers, this approach also promises to allow a fair comparison if the raw prevalence $\pi_c$ is compared with $e_c$. Typically, the ratio $\pi_c/e_c$ (often referred to as O/E, or standardized morbidity ratio) is used to quantify the difference, but considering their difference has been proposed too.[6,23,25,28]

There are different ways to define the reference population. We consider in this article 2 approaches. In general, if $q(x)$ describes the relationship between the outcome and the covariates in the reference population, and a corresponding estimate $\hat{q}$ is available, then we can formally consider that

$$e_c = E(q(X) \mid C = c)$$

is approximated by

$$\hat{e}_c = \frac{1}{n_c} \sum_{i, C_i = c} \hat{q}(X_i),$$

with $X_i$ denoting the covariate values of patient $i$ among the $n_c$ patients from center $c$.

In the first approach, the whole patient population is used as the reference population and $q(x)$ is equated with

$$p(x) = P(Y = 1 \mid X = x).$$

Typically, estimation of $p(x)$ is based on a logistic regression model. Such a model can be represented as

$$\operatorname{logit} p(x) = g_\beta(x).$$

In contrast to the conditional model (cf. section 2.1), the center effects are not included in this model. In the statistical literature, this type of model is called a marginal model (because there is no conditioning on $C = c$), and hence we denote the regression parameter by $\beta^M$.

In the second approach, the population of an "average center" is used as the reference population. The latter is defined based on the conditional model introduced in Section 2.1 as a center with $\xi_c = 0$. Writing $p_c(x)$ as $p(\xi_c, x)$, then $q(x)$ is equated with $p(0, x)$.

Note that $q(x)$ has a very similar structure in the 2 approaches. We have $\operatorname{logit} q(x) = g_{\beta^M}(x)$ in the first approach and $\operatorname{logit} q(x) = g_{\beta^C}(x)$ in the second approach. Hence, from a computational perspective, the only difference is the use of 2 different estimates of $\beta$.

From a historical perspective, the first approach continues the tradition of indirect standardization dating back to the 19th century. In this tradition, the covariate of interest was typically age, divided into age groups, and age-group-specific outcome rates were combined in a weighted manner. This procedure coincides with the first approach described above when the age groups enter the model as a categorical covariate. The second approach became popular when fitting hierarchical logistic regression models became computationally feasible around the end of the last century.[1,6] The 2 approaches have been compared in several case studies[6,13,14,29–31] and included in systematic investigations.[25,27] Most of these investigations have reported only small differences if the conditional model is estimated via a fixed-effects approach.

Further variants of indirect standardization can be found in the literature,[18] which additionally replace $\pi_c$ by model-based quantities. These variants are not considered in this article.

# INDIRECT STANDARDIZATION BASED ON AN AUGMENTED MARGINAL MODEL

For pedagogical reasons, we now introduce an additional variant of indirect standardization based on an augmented marginal model. In this variant, we try to take into account that the mean $\mu_c$ of the patient characteristics for center $c$ may play its own role in determining the average outcome of center $c$.

The idea is to focus on patients from centers with similar mean values when determining the expected outcomes. Formally, this can be based on estimating

$$p(x, \mu) = P(Y = 1 \mid X = x, \mu_c = \mu)$$

and considering

$$e_c^A = E(p(X, \mu_c) \mid C = c).$$

Estimation of $p(x, \mu)$ can be based on the logistic regression model

$$\operatorname{logit} p(x, \mu) = g_{\beta^A}(x) + h_{\gamma^A}(\mu)$$

with $h_{\gamma^A}(\mu)$ denoting a prespecified function of the mean value and some regression parameters $\boldsymbol{\gamma^A}$.

Formally, $e_c^A$ describes the expected prevalence in centers with the same mean value of the covariates. The intended interpretation as the expected prevalence in centers with similar mean values is allowed by noting that the augmented marginal model borrows information from centers with similar mean values.

Such augmented models have been studied in the statistical literature as attempts to estimate simultaneous between- and within-cluster covariate effects.[32,33]

## AN ARTIFICIAL CASE STUDY

In this artificial case study, we consider *readmission* as the outcome and *age* as the only covariate of interest for 21 centers. There are seven different mean ages ($\mu_c = 50$, 55, 60, 65, 70, 75, or 80, respectively), each shared by 3 centers. The association between age and readmission probability at the individual level in center $c$ is defined by the equation

$$\operatorname{logit} p_c(age) = -\,3.4 + \xi_c + \beta^C age,$$

with $\beta^C = 0.028$. This means that increasing age by 10 years within a center multiplies the odds of readmission by exp(0.028 × 10) = 1.32. The center effects $\xi_c$ are chosen as +0.25, 0, and −0.25, within each group of 3 centers sharing the same mean age. The whole setup is depicted in Figure 1 (scenario A). Centers with $\xi_c = +0.25$, 0, −0.25, respectively, can be regarded as performing below average (highest readmission rate when comparing centers with the same mean age), on average, and above average (lowest readmission rates when comparing centers with the same mean age), respectively (cf. Fig. 1). We consider the risk adjustment methods to be doing their job well if they allow us to distinguish these 3 groups of centers.

Now let us assume that this setup describes the situation at a specific time point. From then on, centers with higher mean age may start to make more efforts to reduce the readmission rate than centers with lower mean age. This may have happened because they experienced a rather high readmission rate compared with other centers (with younger patients on average) or because they experienced their high readmission rate as a crucial issue for their patients or for the economic situation of the center. If these efforts are successful, the readmission rates may decrease in all centers, but the amount of the decrease increases with mean age.

This leads to the definition of a second scenario (scenario B) which tries to mimic the situation at the end of such a process. It differs from scenario A only with respect to the center effects. The center effects in scenario B are related to the center effects in Scenario A via

$$\xi_c^B = \xi_c^A - 0.01 \times \left(\mu_c - 50\right).$$

This choice implies no difference in the center effects between scenario A and scenario B if the mean age is 50, and an increasing reduction of the center effects with increasing mean age. This reflects the differential improvement in the quality of healthcare between centers with patients of different mean age. Therefore, in scenario B, the center effects are correlated with the mean age of the patients. In scenario A, this was not the case: the set of center effects was identical for each possible mean age. Figure 1 also depicts scenario B to allow a comparison with scenario A.

As we apply the same reduction to any set of 3 centers sharing the same mean age, the relative position of the 3 centers and the differences between the 3 centers (on the logit scale) remain the same. The arguments presented above with regard to the centers being above, on, or below average still apply under scenario B. Consequently, it seems reasonable to expect a risk adjustment method to do the job of distinguishing the 3 groups defined above, irrespective of their mean age.

It might be argued that it is unfair that centers with the highest mean age do not get a reward for reaching a higher quality improvement than centers with (on average) younger patients. However, this is only apparent when knowing the process of reaching scenario B from scenario A. In practice, provider comparisons are based only on data from one time period. If scenario B provides the available information, then that process is unknown.

Figure 2 depicts the results we can expect when applying the different methods to both scenarios, assuming a very large number of patients in each center, allowing us to neglect statistical uncertainty (details of the computations behind Fig. 2 are outlined in Supplemental Digital Content 1, Appendix 1, http://links.lww.com/MLR/C838. The corresponding code is documented in Supplemental Digital Content 2, Appendix 2, http://links.lww.com/MLR/C839). Under scenario A, all methods give very similar results. The 3 groups of centers are clearly discriminated, and within each group there is little variation. The variation within each group can be explained by the fact that center effects in the data-generating model refer to differences on the logit scale, whereas all methods perform a transformation to the probability scale. This is an example of a "negligible" systematic difference, as mentioned in the introduction.

The picture is completely different under scenario B. Only indirect standardization based on the augmented marginal model reflects clearly that quality differences are related in the first place to the 3 groups. For all other methods, the values obtained reflect both the differences between the 3 groups and the differences in the mean age across the centers. In the case of direct standardization and indirect standardization based on the conditional model, the influence of the latter is so strong that the values computed within each group overlap across groups. For example, this implies that, even after risk adjustment, some below-average centers obtain more favorable (i.e., lower values) than some on-average centers. Arguably, risk adjustment has not done its job well here.

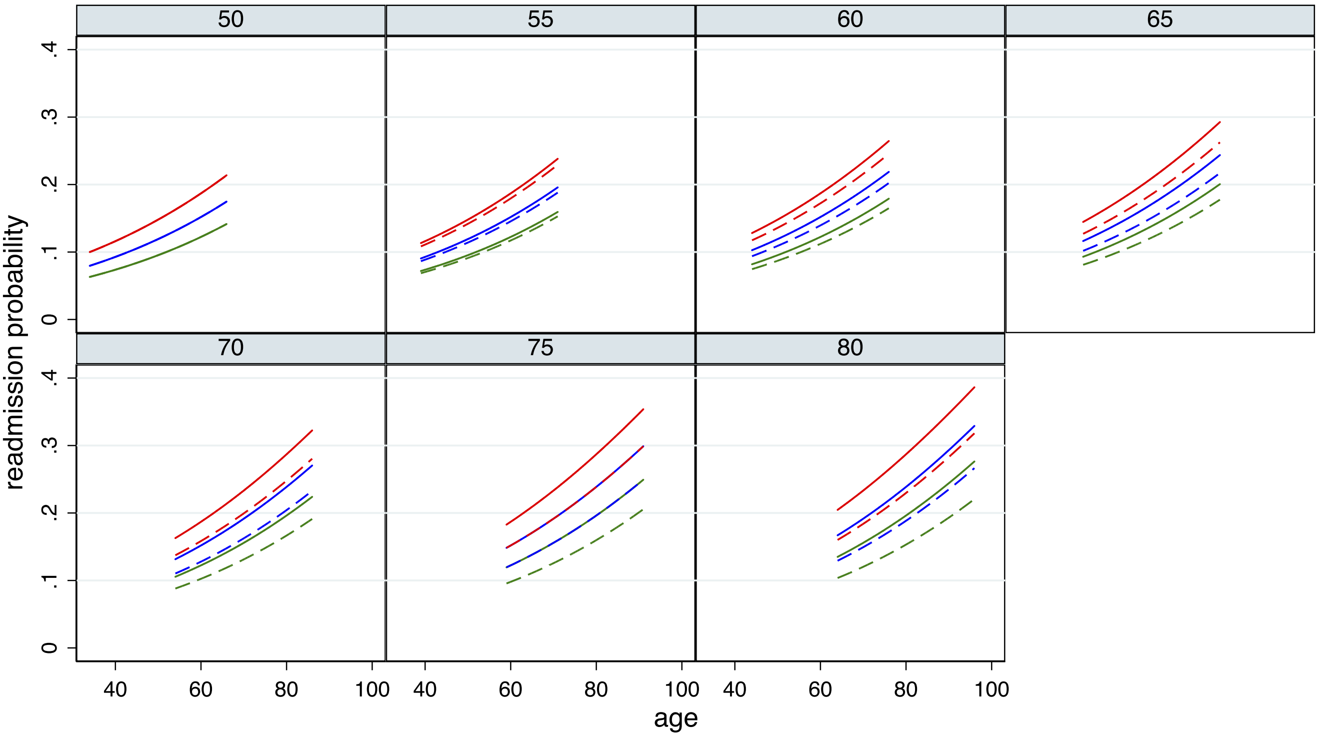

**FIGURE 1.** The functions $p_c(x)$ for the 21 centers for scenario A (solid lines) and scenario B (dashed lines). The 7 panels divide the centers according to their mean age. The functions are drawn over a range of age values from the 5th to the 95th percentile of the distribution of age within each center. The color reflects the magnitude of the center effects in scenario A (0.25: red; 0: blue; −0.25: green).

## EXPLAINING THE OBSERVED RESULTS

The results obtained with direct standardization for scenario B are rather easy to explain. In both scenarios, the center effect estimates of the conditional model are estimates of the center effects in the data-generating model. This is to be expected from statistical theory and it is corroborated by the numerical results presented in Table 1. In scenario A, the center effects reflect the 3 center groups. In scenario B, the center effects also reflect the mean age of the centers. As mentioned above, the values $d_c$ obtained by direct standardization represent a monotone transformation of the center effects from the conditional model. Consequently, the ranking of these values partially reflects the differences in mean age—as observed from Figure 2.

The results obtained by indirect standardization for Scenario B can be explained by considering the values $e_c$ obtained by using the 3 variants in relation to the raw prevalence values (Fig. 3) and the values of the regression parameters for the 3 models involved (Table 1). When using the augmented marginal model, the definition of $e_c^A$ takes into account that the probability of readmission for an individual depends both on their age and the mean age of their center. The regression coefficients of the augmented marginal model indicate an increase with individual age ($\beta^A(age) = 0.0278$) and a decrease with the mean age of the center ($\beta^A(\mu(age)) = -0.0099$). This results in values $e_c^A$ which are close to the raw prevalence values for the center in the middle group, which agrees perfectly with the intended role of $e_c^A$: it should reflect the expected outcome over all centers with the same mean age.

When using the marginal model, the influence of the center-specific mean age on the outcome of an individual is ignored. We can observe a partial compensation for the absence of the negative value of $\beta^A(\mu(age))$: $\beta^M(age)$ is smaller than $\beta^A(age)$. However, the difference $\beta^M(age) - \beta^A(age) = -0.005$ is only half of $\beta^A(\mu(age)) = -0.01$. Consequently, $e_c^M$ is larger than $e_c^A$ if the mean age is large, $e_c^M$ is smaller than $e_c^A$ if the mean age is small, and then the ratios O/E partially reflect the differences in mean age.

When using the conditional model, $\beta^C(age)$ is close to $\beta^A(age)$. This reflects the fact that the conditional model uses only the association between age and readmission within each center to assess the age effect – everything else is covered by the center effects. As $\beta^C(age)$ is larger than $\beta^M(age)$, the association of the ratios O/E with mean age becomes more pronounced.

Note that under Scenario A, the values of $\beta^A(age)$, $\beta^M(age)$, and $\beta^C(age)$ only differ slightly and $\beta^A(\mu(age))$ is close to 0. Consequently, the 3 versions of indirect standardization give very similar results. With respect to the 2 versions described in Section 2.2., this is in line with the empirical case studies mentioned above[6,13,14,29–31] reporting only small differences between the 2 versions. The slight differences between $\beta^M(age)$ and $\beta^C(age)$ reflect

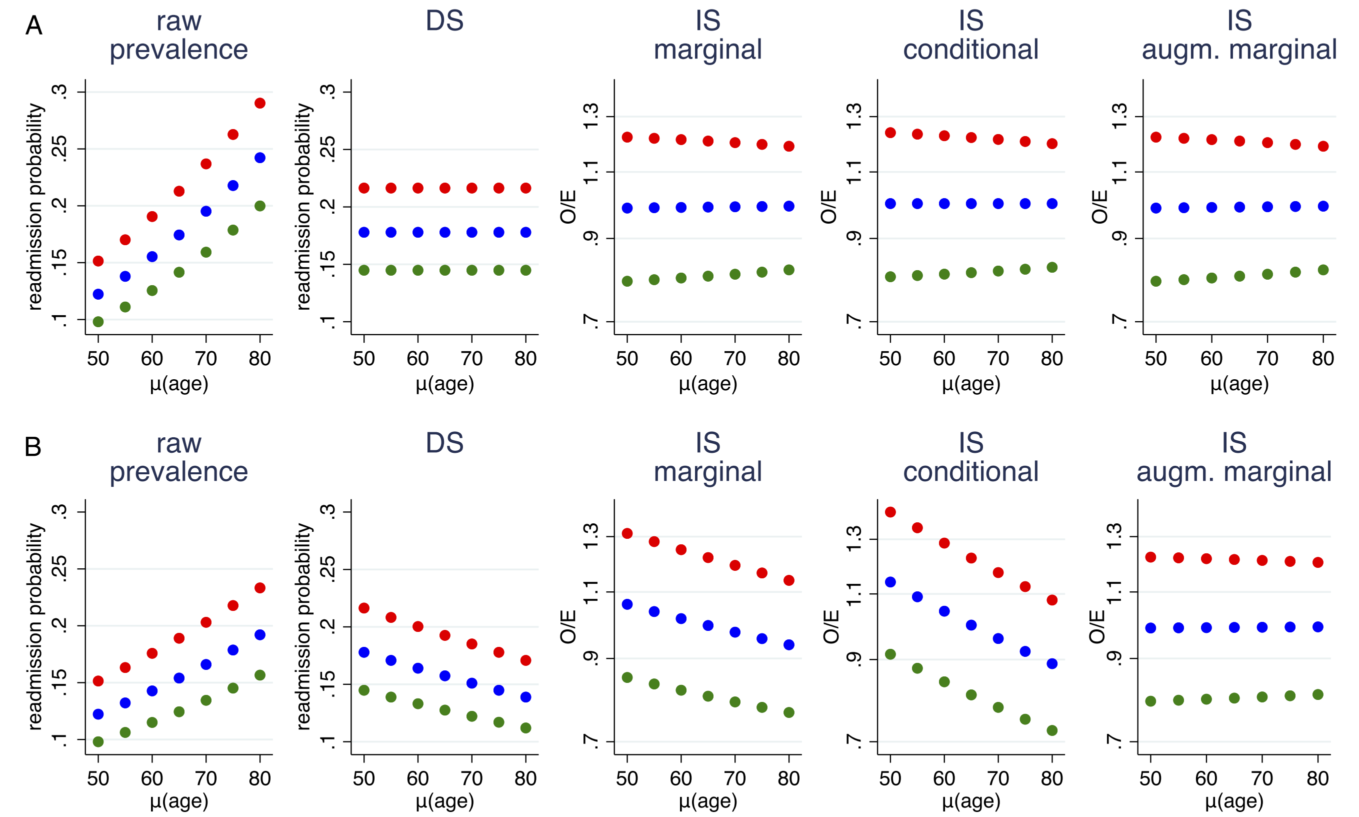

**FIGURE 2.** The raw prevalence values and the results of four risk adjustment methods for scenarios A and B, assuming very large sample sizes in each center. The centers are arranged according to their mean age. The colors are chosen as in Figure 1. Details of the computations are outlined in Supplemental Digital Content 2, Appendix 1 (http://links.lww.com/MLR/C838). The corresponding code is documented in Supplemental Digital Content 2, Appendix 2 (http://links.lww.com/MLR/C839). DS indicates direct standardization; IS, indirect standardization.

a well-known effect of adding to a logistic regression model a covariate which is independent of the other covariates.[34] This is another example of a systematic difference of "negligible" magnitude as mentioned in the Introduction.

## DISCUSSION

In this article, we tried to shed some further light on the impact of an association between the center effects and the distributions of patient characteristics in the centers on provider comparisons. We approached this by considering a single artificial case study with center effects correlated with the center-specific mean values of a patient characteristic.

Generalizing from a single case study is always challenging. However, we believe that it is fair to draw the following general conclusions: If center effects in a conditional logistic model are correlated with the center-specific mean values of a patient characteristic, then

- indirect standardization based on a marginal model may give systematically different results compared to indirect standardization based on a conditional model.
- center effects do not reflect deviations from the average, but also the correlation of center effects with the mean values.
- even if we can estimate the center effects in an unbiased manner, they cannot be used directly to compare provider performance.
- direct standardization is at risk of producing misleading results, as the corresponding values are monotone transformations of the center effects.
- the traditional risk adjustment methods considered may not do their job well, as they are at risk to reflect differences in the mean values across the centers while masking other differences of higher interest.

These findings have to be contrasted with the current practice of provider comparisons, which is characterized by a lack of consensus about the choice of the risk adjustment method—the choice is typically based on some institutional standard or is at the discretion of a research group. This lack of consensus may reflect an implicit consensus about a simple rule: All risk adjustment methods give usually very similar results. This coincides with our findings for Scenario A. However, Scenario B suggests that there are exceptions to this rule.

**TABLE 1.** The Parameter Values of the Data-Generating Model and of the 3 Models Used for Risk Adjustment

| Scenario A | | | | | | |
|---|---|---|---|---|---|---|
| **Parameter** | ***c*** | **$\mu_c$** | **Data-Generating Model** | **Marginal Model** | **Conditional Model** | **Augmented Marginal Model** |
| $\xi_c$ | 1 | 50 | −0.2500 | — | −0.2500 | — |
| | 2 | 50 | 0.0000 | — | −0.0000 | — |
| | 3 | 50 | 0.2500 | — | 0.2500 | — |
| | 4 | 55 | −0.2500 | — | −0.2500 | — |
| | 5 | 55 | 0.0000 | — | 0.0000 | — |
| | 6 | 55 | 0.2500 | — | 0.2500 | — |
| | 7 | 60 | −0.2500 | — | −0.2500 | — |
| | 8 | 60 | 0.0000 | — | 0.0000 | — |
| | 9 | 60 | 0.2500 | — | 0.2500 | — |
| | 10 | 65 | −0.2500 | — | −0.2500 | — |
| | 11 | 65 | 0.0000 | — | −0.0000 | — |
| | 12 | 65 | 0.2500 | — | 0.2500 | — |
| | 13 | 70 | −0.2500 | — | −0.2500 | — |
| | 14 | 70 | 0.0000 | — | 0.0000 | — |
| | 15 | 70 | 0.2500 | — | 0.2500 | — |
| | 16 | 75 | −0.2500 | — | −0.2500 | — |
| | 17 | 75 | 0.0000 | — | −0.0000 | — |
| | 18 | 75 | 0.2500 | — | 0.2500 | — |
| | 19 | 80 | −0.2500 | — | −0.2500 | — |
| | 20 | 80 | 0.0000 | — | 0.0000 | — |
| | 21 | 80 | 0.2500 | — | 0.2500 | — |
| $\beta^M(age)$ | — | — | — | 0.0278 | — | — |
| $\beta^C(age)$ | — | — | 0.0280 | — | 0.0280 | — |
| $\beta^A(age)$ | — | — | — | — | — | 0.0278 |
| $\beta^A(\mu(age))$ | — | — | — | — | — | 0.0000 |
| Scenario B | | | | | | |
| parameter | *c* | $\mu_c$ | data generating model | marginal model | conditional model* | augmented marginal model |
| $\xi_c$ | 1 | 50 | −0.2500 | — | −0.1000 | — |
| | 2 | 50 | 0.0000 | — | 0.1500 | — |
| | 3 | 50 | 0.2500 | — | 0.4000 | — |
| | 4 | 55 | −0.3000 | — | −0.1500 | — |
| | 5 | 55 | −0.0500 | — | 0.1000 | — |
| | 6 | 55 | 0.2000 | — | 0.3500 | — |
| | 7 | 60 | −0.3500 | — | −0.2000 | — |
| | 8 | 60 | −0.1000 | — | 0.0500 | — |
| | 9 | 60 | 0.1500 | — | 0.3000 | — |
| | 10 | 65 | −0.4000 | — | −0.2500 | — |
| | 11 | 65 | −0.1500 | — | 0.0000 | — |
| | 12 | 65 | 0.1000 | — | 0.2500 | — |
| | 13 | 70 | −0.4500 | — | −0.3000 | — |
| | 14 | 70 | −0.2000 | — | −0.0500 | — |
| | 15 | 70 | 0.0500 | — | 0.2000 | — |
| | 16 | 75 | −0.5000 | — | −0.3500 | — |
| | 17 | 75 | −0.2500 | — | −0.1000 | — |
| | 18 | 75 | 0.0000 | — | 0.1500 | — |
| | 19 | 80 | −0.5500 | — | −0.4000 | — |
| | 20 | 80 | −0.3000 | — | −0.1500 | — |
| | 21 | 80 | −0.0500 | — | 0.1000 | — |
| $\beta^M(age)$ | — | — | — | 0.0228 | — | — |
| $\beta^C(age)$ | — | — | 0.0280 | — | 0.0280 | — |
| $\beta^A(age)$ | — | — | — | — | — | 0.0278 |
| $\beta^A(\mu(age))$ | — | — | — | — | — | −0.0099 |

c: The index of the 21 centers. $\mu_c$: The mean age of center c. $\xi_c$: The intercept of center c in the data generating model. $\beta^M$(age): The regression coefficient of age in a marginal logistic regression model. $\beta^C$(age): The regression coefficient of age in a conditional logistic regression model. $\beta^A$(age): The regression coefficient of age in an augmented logistic regression model. $\beta^A$(μ(age)): The regression coefficient of the center specific mean age in an augmented logistic regression model.

*In Scenario B there is a constant difference of 0.15 between the values from the data-generating model and the values from the conditional model, as only the latter values are centered at 0.

This calls for several actions. First, for any (new) specific application of risk adjustment methods, it seems to be wise to check the sensitivity of the results to the choice of the risk adjustment method. Second, existing applications should be reanalyzed retrospectively in order to estimate the likelihood of such sensitivity and to identify conditions increasing the risk of such sensitivity. Third, we should develop strategies to incorporate information from longitudinal data in provider comparisons and apply them retrospectively in order to understand their merits. Finally, there is interest in developing risk adjustment methods that are doing their job also for scenarios characterized by

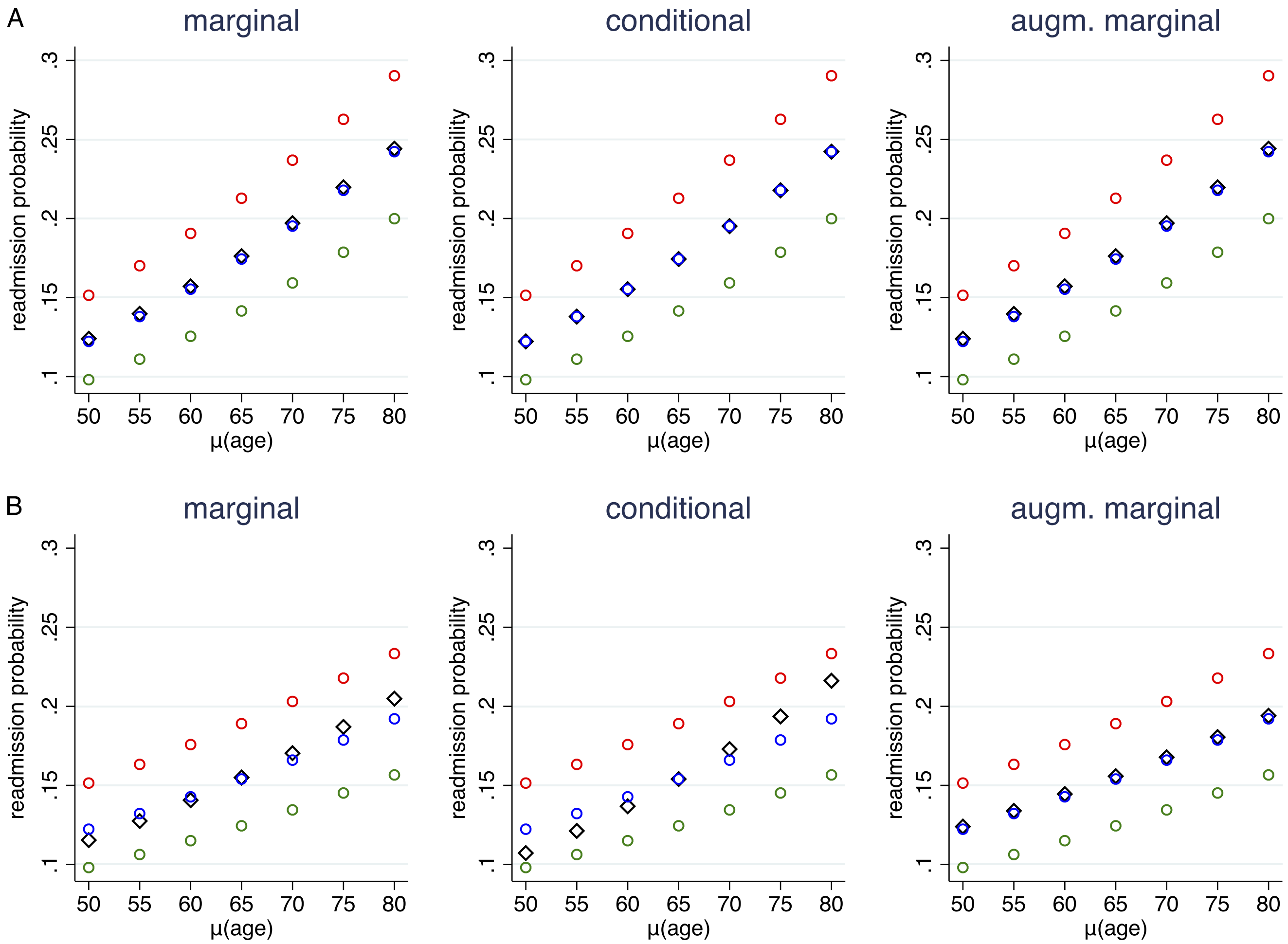

**FIGURE 3.** The values of $e_c$ for each center obtained by the 3 variants of indirect standardization for both scenarios (black diamonds). These values are identical for centers with the same mean age. In addition, the raw prevalence value of each center is shown by a dot. The colors of the dots are chosen as in Figure 1. The centers are arranged according to their mean age.

correlations between center effects and the center-specific mean values of patient characteristics. To which degree the augmented marginal model (or similar models) can play a role has to be clarified. In Supplementary Appendix 3, Supplemental Digital Content 3, http://links.lww.com/MLR/C840 we elaborate on all four actions.

Three articles have previously considered the impact of a correlation between center effects and the center-specific mean values of covariates by use of finite sample simulation studies. Kalbfleisch and Wolfe[15] observed that fixed effect estimates of center effects remain valid, which coincides with our observations in Table 1. Roessler, Schmitt and Schoffer[25] studied the impact of several variants of indirect standardization on the correct classification of centers. Varewyck et al[21] considered a more complex situation by also allowing patient characteristics to change the effect of covariates within each center. In all these articles, however, the center effects of the data-generating model served as ground truth, and as such their results are not comparable with our results.

The failure of direct standardization to discriminate between the 3 groups of centers in Scenario B may reflect a deeper, conceptual problem. Direct standardization assumes that the relationship between patient characteristics and outcome observed in one center can be applied to all patients, and this mimics the situation that all patients would have been treated at this center. However, if characteristics of the patient population change the performance of the center (as in our artificial case study, where centers improved their quality in response to their specific patient population), this rationale breaks down, as the whole population typically differs with regard to the distribution of patient characteristics from the population of a single center. Indirect standardization assumes instead that the relationship observed in some reference population can be applied in the center of interest. However, our investigation suggests that it might not be always adequate to use all patients as the reference population.

A general challenge highlighted by our article is the need for new frameworks to answer the simple question of whether risk adjustment methods do their job well in this specific context. Many investigations about risk adjustment models start with considering the conditional model as the data-generating model with center effects drawn at random (and independently from any other information). In that case, common risk adjustment methods tend to generate values that are monotone transformations of the center effects, and hence the question about how to define the truth does not arise. This is no longer the case if center effects are associated with the distributions of patient characteristics, a situation which cannot be excluded in practice. We presented an alternative approach by defining a scenario in which there might be a consensus about what we should expect from risk adjustment methods with respect to ranking of the centers. This seems to us to be a fruitful approach, and the development of further scenarios is desirable.

Finally, the artificial case study considered in this article may question the use of the general approach of basing provider comparisons on a cross sectional, annual assessment. The definition of a fair comparison may need to take into account quality improvements observable in longitudinal data.

## ACKNOWLEDGMENTS

*The authors are grateful to Jona Cederbaum and Stefan Gehrig for their constructive comments.*

## REFERENCES

1. Normand SLT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc*. 1997;92:803–814.
2. Jarman B, Gault S, Alves B, et al. Explaining differences in English hospital death rates using routinely collected data. *Brit Med J*. 1999; 318:1515–1520.
3. Miyata H, Hashimoto H, Horiguchi H, et al. Performance of in-hospital mortality prediction models for acute hospitalization: hospital standardized mortality ratio in Japan. *BMC Health Serv Res*. 2008;8:229.
4. Busch P, Fehr N. A model with a future: standardised quality reviews in Swiss hospitals and clinics. *Hospital Management in Europe: Official Journal of the European Association of Hospital Managers*. 2021;2:10–11.
5. Berwick DM, James B, Coye MJ. Connections between quality measurement and improvement. *Med Care*. 2003;41(1 Suppl):I30–I38.
6. DeLong ER, Peterson ED, DeLong DM, et al. Comparing risk-adjustment methods for provider profiling. *Stat Med*. 1997;16: 2645–2664.
7. Ash AS, Ellis RP. Risk-adjusted payment and performance assessment for primary care. *Med Care*. 2012;50:643–653.
8. Normand SLT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. *Stat Sci*. 2007;22:206–226.
9. O'Hara LM, Masnick M, Leekha S, et al. Indirect vs. direct standardization methods for reporting healthcare-associated infections: an analysis of central line-associated bloodstream infections in Maryland. *Infect Control Hosp Epidemiol*. 2017;38:989–992.
10. He K. Indirect and direct standardization for evaluating transplant centers. *J Hosp Adm*. 2018;8:9.
11. Rixom A. Performance league tables. *Brit Med J*. 2002;325:177–178.
12. Higham DJH, Dr Julian Flowers DDP Direct standardisation has no advantage over indirect standardisation. Accessed December 3, 2022. https://www.bmj.com/rapid-response/2011/10/29/direct-standardisation-has-no-advantage-over-indirect-standardisation
13. Li Y, Cai X, Glance LG, et al. National release of the nursing home quality report cards: implications of statistical methodology for risk adjustment. *Health Serv Res*. 2009;44:79–102.
14. Alexandrescu R, Jen MH, Bottle A, et al. Logistic versus hierarchical modeling: an analysis of a statewide inpatient sample. *J Am Coll Surg*. 2011;213:392–401.
15. Kalbfleisch JD, Wolfe RA. On monitoring outcomes of medical providers. *Stat Biosci*. 2013;5:286–302.
16. Yang X, Peng B, Chen R, et al. Statistical profiling methods with hierarchical logistic regression for healthcare providers with binary outcomes. *J Appl Stat*. 2014;41:46–59.
17. Glance LG, Li Y, Dick AW. Quality of quality measurement: impact of risk adjustment, hospital volume, and hospital performance. *Anesthesiology*. 2016;125:1092–1102.
18. Mohammed MA, Manktelow BN, Hofer TP. Comparison of four methods for deriving hospital standardised mortality ratios from a single hierarchical logistic regression model. *Stat Methods Med Res*. 2016;25:706–715.
19. Şentürk D, Chen Y, Estes JP, et al. Impact of case-mix measurement error on estimation and inference in profiling of health care providers. *Commun Stat Simul Comput*. 2020;49:2206–2224.
20. Mu Y, Chin AI, Kshirsagar AV, et al. Assessing the impacts of misclassified case-mix factors on health care provider profiling: performance of dialysis facilities. *Inquiry*. 2020;57: 0046958020919275.
21. Varewyck M, Vansteelandt S, Eriksson M, et al. On the practice of ignoring center-patient interactions in evaluating hospital performance. *Stat Med*. 2016;35:227–238.
22. Moran JL, Santamaria JD, Duke GJThe Australian & New Zealand Intensive Care Society (ANZICS) Centre for Outcomes & Resource Evaluation (CORE). Modelling hospital outcome: problems with endogeneity. *BMC Med Res Methodol*. 2021;21:124.
23. Varewyck M, Goetghebeur E, Eriksson M, et al. On shrinkage and model extrapolation in the evaluation of clinical center performance. *Biostatistics*. 2014;15:651–664.
24. Daignault K, Saarela O. Doubly robust estimator for indirectly standardized mortality ratios. *Epidemiologic Methods*. 2017;6: 20160016.
25. Roessler M. Schmitt JSchoffer O. Ranking hospitals when performance and risk factors are correlated: a simulation-based comparison of risk adjustment approaches for binary outcomes. *PLoS One*. 2019;14:e0225844.
26. Keiding N, Clayton D. Standardization and control for confounding in observational studies: a historical perspective. *Stat Sci*. 2014;29: 529–558.
27. Kristoffersen DT, Helgeland J, Clench-Aas J, et al. Observed to expected or logistic regression to identify hospitals with high or low 30-day mortality? *PLoS ONE*. 2018;13:e0195248.
28. Goetghebeur E, Van Rossem R, Baert K, et al. *Quality Inssurance of rectal cancer diagnosis and treatment – phase 3: statistical methods to benchmark centers on a set of quality indicators*. Brussels: Belgian Health Care Knowledge Centre (KCE); 2011. (KCE Reports). Report No.: 161C.
29. Fedeli U, Brocco S, Alba N, et al. The choice between different statistical approaches to risk-adjustment influenced the identification of outliers. *J Clin Epidemiol*. 2007;60:858–862.
30. Alexandrescu R, Bottle A, Jarman B, et al. Classifying hospitals as mortality outliers: logistic versus hierarchical logistic models. *J Med Syst*. 2014;38:29.
31. Glance LG, Dick A, Osler TM, et al. Impact of changing the statistical methodology on hospital and surgeon ranking: the case of the New York State cardiac surgery report card. *Med Care*. 2006;44: 311–319.
32. Neuhaus JM, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*. 1998;54:638.
33. Berlin JA, Kimmel SE, Ten Have TR, et al. An empirical comparison of several clustered data approaches under confounding due to cluster effects in the analysis of complications of coronary angioplasty. *Biometrics*. 1999;55:470–476.
34. Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*. 1993;80: 807–815.