Research Article

Statistics in Medicine

Received 13 January 2015, Accepted 7 August 2015 Published online 24 August 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6634

On the practice of ignoring center-patient interactions in evaluating hospital performance

Machteld Varewyck,[a*†] Stijn Vansteelandt,[a] Marie Eriksson[b] and Els Goetghebeur[a]

We evaluate the performance of medical centers based on a continuous or binary patient outcome (e.g., 30-day mortality). Common practice adjusts for differences in patient mix through outcome regression models, which include patient-specific baseline covariates (e.g., age and disease stage) besides center effects. Because a large number of centers may need to be evaluated, the typical model postulates that the effect of a center on outcome is constant over patient characteristics. This may be violated, for example, when some centers are specialized in children or geriatric patients. Including interactions between certain patient characteristics and the many fixed center effects in the model increases the risk for overfitting, however, and could imply a loss of power for detecting centers with deviating mortality. Therefore, we assess how the common practice of ignoring such interactions impacts the bias and precision of directly and indirectly standardized risks. The reassuring conclusion is that the common practice of working with the main effects of a center has minor impact on hospital evaluation, unless some centers actually perform substantially better on a specific group of patients and there is strong confounding through the corresponding patient characteristic. The bias is then driven by an interplay of the relative center size, the overlap between covariate distributions, and the magnitude of the interaction effect. Interestingly, the bias on indirectly standardized risks is smaller than on directly standardized risks. We illustrate our findings by simulation and in an analysis of 30-day mortality on Riksstroke. © 2015 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Keywords: causal effects; direct and indirect standardization; Firth correction; misspecified model; quality of care

1. Introduction

Many continuing efforts are made to improve the accuracy of hospital quality of care assessments [1, 2]. They are motivated by the major impact of performance evaluations not only on the improvement of care but also on the patient's choice of hospital, or financial pay-per-performance incentives, for example. Key aspects of the quality of hospital performance are commonly evaluated through a binary or continuous outcome quality indicator via direct or indirect standardization [3, 4]. Direct standardization aims to assess for each center how the entire study population would have fared under its current level of care. Indirect standardization contrasts the quality outcome in each center with what is expected should their patients choose randomly over the level of care across all centers. Because the choice of standardization technique depends on the research question, we will report results for both techniques. Traditionally, indirect standardization is used, and this is most relevant to judge center performance when the center's own population does not substantially change over time. Should centers vary in approach and one wishes

[a]*Department of Applied Mathematics, Computer Science and Statistics, Ghent University, 9000 Ghent, Belgium*
[b]*Department of Statistics, Umeå University, 901 87 Umeå, Sweden*
[*]*Correspondence to: Machteld Varewyck, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, 9000 Ghent, Belgium.*
[†]*E-mail: machteld.varewyck@ugent.be*
The copyright line for this article was changed on 10 September 2015 after original online publication.

to choose one approach for implementation across all centers, then direct standardization delivers the most relevant impact measure.

In general, standardized risks are obtained via outcome regression models that adjust for confounding of the center-outcome effect [5, 6]. Adjustment for differential patient mix is necessary, because centers treating for instance older patients will show a higher mortality risk irrespective of their actual treatment quality. Interactions between center and patient characteristics are rarely modelled, however. This is due to the curse of dimensionality, which may already show up in the main effects model. One may hit low information content to estimate the main effects of many centers, resulting in large finite sample bias. For this reason, we have chosen to use Firth-corrected fixed effects regression instead of standard fixed effects regression [7]. This method is preferred over standard random effects regression, which may unduly shrink center effects toward the overall mean [1], thereby masking outlying performance especially of the smallest centers [7].

In practice, center effects may interact with patient characteristics when some centers perform particularly well on a specific subgroup of patients [2, 8, 9]. In [10] for instance, differences in estimated 30-day mortality risk between hospitals were relatively small for patients with low illness severity but variation increased for patients with increasing illness severity. This may indicate that high-risk patients receive more specialized care at some hospitals. Similarly in [1], the variation in hospital effect was substantial and depended on patient baseline severity. In a next step, it is important to know which hospitals and why they show interactions. Evidence suggested that effect differences were sometimes associated with hospital size, urbanicity, and academic affiliation, as medium-sized hospitals had slightly weaker effects of baseline severity on the outcome than large hospitals.

In this article, we will therefore study to what extent bias may enter each of the standardized risks when modeling constant center effects across patient profiles while interactions between center and patient characteristics are present. This may justify the common practice of ignoring center-patient interactions, especially in situations where it is simply prohibitive to allow for effect modification because sufficient information is lacking in small centers, for example. In [11], fitting problems are overcome by modeling interactions between patient characteristics and hospital type (teaching status, location), but this limits evaluations to hospital groups rather than individual hospitals.

2. Setting

Throughout the paper, C will denote a random variable indicating in which center the patient was actually treated ($C = 1, \ldots, m$) and $\mathbf{L}$ denotes the vector of patient-specific baseline characteristics such as gender, age, and initial disease status. We focus on the following data-generating outcome regression model:

$$E(Y|\mathbf{L}, C) = g\left(\sum_{c=1}^{m} I(C = c)(\mathbf{L}'\boldsymbol{\beta}_c + \psi_c)\right), \tag{1}$$

which allows center effects to depend on $\mathbf{L}$. So, the expected outcome in center c is parameterized by ψ_c for a patient with the reference profile ($\mathbf{L} = \mathbf{0}$) and by $\psi_c + \mathbf{l}_0'\boldsymbol{\beta}_c$ for a patient with $\mathbf{L} = \mathbf{l}_0$ profile. Here, $g(\cdot)$ is a known link function, for example, the logistic link.

2.1. *Nature of interactions*

In the outcome regression model (1), interactions between center and patient-specific characteristics may arise in different ways. We are interested in the case where some centers perform structurally better on a specific subgroup. For example, the difference in care between hospitals is not constant among age groups when younger patients acquire very similar care in each center while older patients receive much better care in some centers compared with others, perhaps because of special equipment or experience of the hospital staff with geriatric patients.

In the absence of such structural interactions, center-patient interactions could still occur due to the scale of the fitted model. While center effects may not interact with patient characteristics on the scale of the linear predictor in a logistic regression model, an interaction may be needed if an additive linear model is used instead [12]. Interactions may also manifest themselves as a result of unmeasured confounding, for example, due to unknown environmental factors. For example, pollution may increase mortality risk in some regions. If the performance of each center is constant over age but the pollution especially affects older patients, then it will induce poorer center effects for older patients in polluted regions. Unmeasured

confounders may thus introduce or hide interactions between center and measured confounders [13]. Throughout, we will exclude this possibility as we will assume that there are no unmeasured confounders, that is

$$Y(c) \perp\!\!\!\perp C|\mathbf{L} \text{ for all } c, \tag{2}$$

where $Y(c)$ indicates the potential outcome for a given patient if he/she were treated at the care level of center c [14].

2.2. Direct and indirect standardization

We will assess the impact on the directly and indirectly standardized risk when interactions between center and patient characteristics are ignored.

Direct standardization aims to infer the potential full population risk for each center c: the risk that would be realized if all patients under study were to experience the care level of that given center c, irrespective of where they were actually treated. We denote this by $E\{Y(c)\}$. Under the outcome regression model in (1) and assuming (2), this can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} g\left(\mathbf{L}_i'\hat{\boldsymbol{\beta}}_c + \hat{\psi}_c\right), \tag{3}$$

for a study population of size n, where $\hat{\boldsymbol{\beta}}_c$ and $\hat{\psi}_c$ are Firth penalized-likelihood estimators [15]. We can then make pairwise comparisons between the directly standardized risk of different centers or with the overall mortality risk $E(Y)$ estimated by

$$\frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{4}$$

In contrast, indirect standardization focuses on what a center achieves for its own patient mix. In general, a risk ratio or risk difference is measured between the observed and expected (e.g., averaged over all care levels) risk in each center [16]. For instance, the excess risk takes the difference between the center's observed risk and the expected risk if its patients were randomly assigned to the care level across the observed distribution of centers, that is

$$\text{Excess risk} = E\{Y(c)|C=c\} - \frac{1}{m}\sum_{c^*=1}^{m} E\{Y(c^*)|C=c\}. \tag{5}$$

Here, the observed risk in center c, $E\{Y(c)|C=c\}$ is estimated by

$$\frac{\sum_{i=1}^{n} Y_i I(C_i=c)}{\sum_{i=1}^{n} I(C_i=c)}. \tag{6}$$

Under the outcome regression model in (1) and assuming (2), the expected risk under the average care level for patients of center c is estimated as follows:

$$\frac{\sum_{i=1}^{n} m^{-1}\sum_{c^*=1}^{m} g\left(\mathbf{L}_i'\hat{\boldsymbol{\beta}}_{c^*} + \hat{\psi}_{c^*}\right) I(C_i=c)}{\sum_{i=1}^{n} I(C_i=c)}. \tag{7}$$

2.3. Ignoring interactions

In the succeeding paragraphs, we will evaluate the bias on estimators (3) and (7) when the working outcome model involves a common center effect γ_c over all patient profiles instead of covariate-specific center effects:

$$E(Y|\mathbf{L},C) = g\left(\mathbf{L}'\boldsymbol{\beta} + \sum_{c=1}^{m} I(C=c)\gamma_c\right). \tag{8}$$

Here, the effect of center c on patient's outcome is expressed by the parameter γ_c, which is now assumed to be the same for each given patient profile.

3. Asymptotic bias calculation

We calculate the asymptotic bias on the directly and indirectly standardized risk when imposing a constant center effect among patients instead of allowing for center-patient interactions in (3) and (5). The average of the observed risks in center c is not model-based when estimated by (6) and therefore unbiased. So, we will calculate the bias on the indirectly standardized risk for center c through the bias on the expected risk when care levels are averaged over all centers, $m^{-1}\sum_{c^*} E\{Y(c^*)|C = c\}$. For simplicity, we focus first on linear regression models including m centers and one patient characteristic L. We fix the number of centers m in our asymptotic calculations, because we focus on the evaluation of centers in a setting where m is relatively fixed (e.g., Riksstroke), but patients come and go. In the Supporting Information, we provide details on the calculations, which are based on a similar principle as in [17].

The asymptotic bias on the directly standardized risk in center c is given by the following:

$$\{E(L|C = c) - E(L)\}\left[\beta_c - \sum_{j=1}^{m} \frac{P(C = j)\mathrm{Var}(L|C = j)}{E\{\mathrm{Var}(L|C)\}}\,\beta_j\right]. \tag{9}$$

For the asymptotic bias on the excess risk for center c we obtain

$$m^{-1}\sum_{c^*=1}^{m}\{E(L|C = c^*) - E(L|C = c)\}\left[\beta_{c^*} - \sum_{j=1}^{m} \frac{P(C = j)\mathrm{Var}(L|C = j)}{E\{\mathrm{Var}(L|C)\}}\,\beta_j\right]. \tag{10}$$

The first factor in these expressions refers to the difference in patient mix between centers, while the second factor contrasts a center-specific L-effect with a weighted average of interaction effects.

Starting with the first factor, it is obvious that for both standardized risks, the bias is zero when all centers have exactly the same L-distribution, so in particular, when L is no confounder of the center-outcome effect; in fact, it suffices that the mean of L is equal in all centers. If not, strong confounding by L implies a small overlap in patient mix between centers or a large 'extrapolation distance' of results from one center to the other, and thus may lead to large bias. The bias increases for a larger deviation of the mean of L in center c from either the mean in the overall population for direct standardization (9) or the mean in any other center for indirect standardization (10). This difference between both standardization techniques can be explained by different extrapolation and is illustrated in Figure 1 for two centers. Direct standardization extrapolates the estimated performance at center c to the whole population under study, while for indirect standardization, the performance of any other center is extrapolated to the patients in center c.

It is clear that the second factor, and thus the bias, is zero when there are no center-patient interactions because then $\beta_1 = \ldots = \beta_m$. We recognize the same weighted sum of interaction effects for both standardization techniques. For a center j, the weight $P(C = j)\mathrm{Var}(L|C = j)E\{\mathrm{Var}(L|C)\}^{-1}$ corresponds to

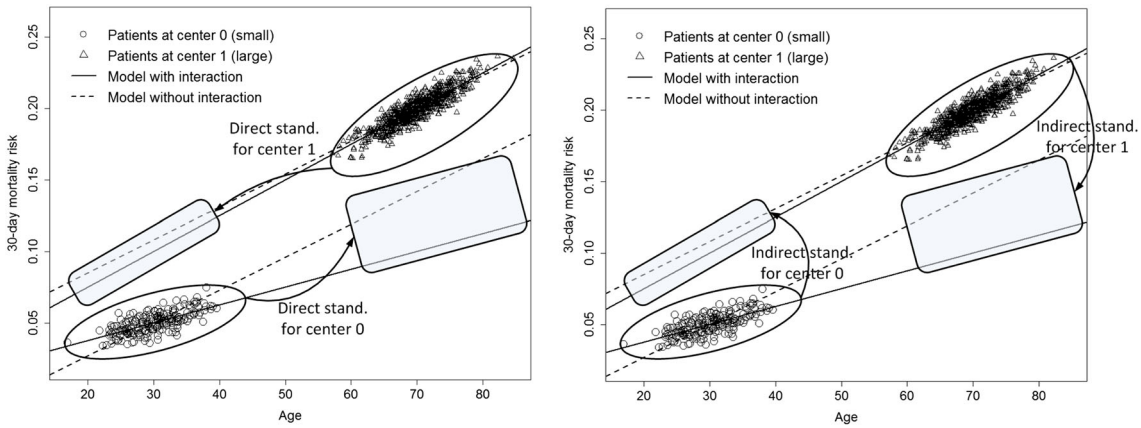


Figure 1. Extrapolation in the directly and indirectly standardized risk considering two centers (small or large center size). The 30-day mortality risk is estimated by a model with or without interaction between center and patient's age.

the relative spread of the patient-mix in that center compared with the other centers, where a larger center size or larger variance of the center-specific L-distribution will result in a larger weight. The weighted sums are then respectively compared with the interaction effect in center c for direct standardization or the interaction effect in any other center than c for indirect standardization. However, in both cases, stronger interactions will result in larger bias.

To obtain more insight in the difference between the bias for direct and indirect standardization, we consider $m = 2$ centers, now coded as $c = 0$ and $c = 1$. Then, the bias on the directly standardized risk for center c (9) simplifies to

$$\{E(L|C = c) - E(L)\}\,(\beta_c - \beta_{1-c})P(C = 1 - c)\,\mathrm{Var}(L|C = 1 - c)\,E\{\mathrm{Var}(L|C)\}^{-1}, \tag{11}$$

and for indirect standardization (10) to

$$\frac{1}{2}\,\{E(L|C = 1 - c) - E(L|C = c)\}\,(\beta_{1-c} - \beta_c)P(C = c)\,\mathrm{Var}(L|C = c)\,E\{\mathrm{Var}(L|C)\}^{-1}. \tag{12}$$

For both standardizations, we again recognize how the difference in average covariate levels and the magnitude of the interaction effect influence the bias. However, they differ in how the center size and the variance of the center-specific L-distribution affect the bias.

Interestingly, for the directly standardized risk, the bias will be larger for relatively small centers, while for the indirectly standardized risk, the bias will be larger for relatively large centers. The different impact of center size for direct and indirect standardization is due to the different extrapolation. The largest center contributes most in estimating the working model parameters, resulting in a smaller bias on the regression line for the largest center (Figure 2, middle panel). Indeed, in Figure 1, the working model does not fit well for the smallest center especially for large values of L (e.g., age) resulting in large bias for the directly standardized risk for that center. For the large center on the other hand, we see little bias on the directly standardized risk. In contrast, for indirect standardization in Figure 1, the smallest center extrapolates to a region where we have good fit, resulting in small bias, while it is the other way around for the large center. For the small center, the expected risk under its own care level is approximately correct

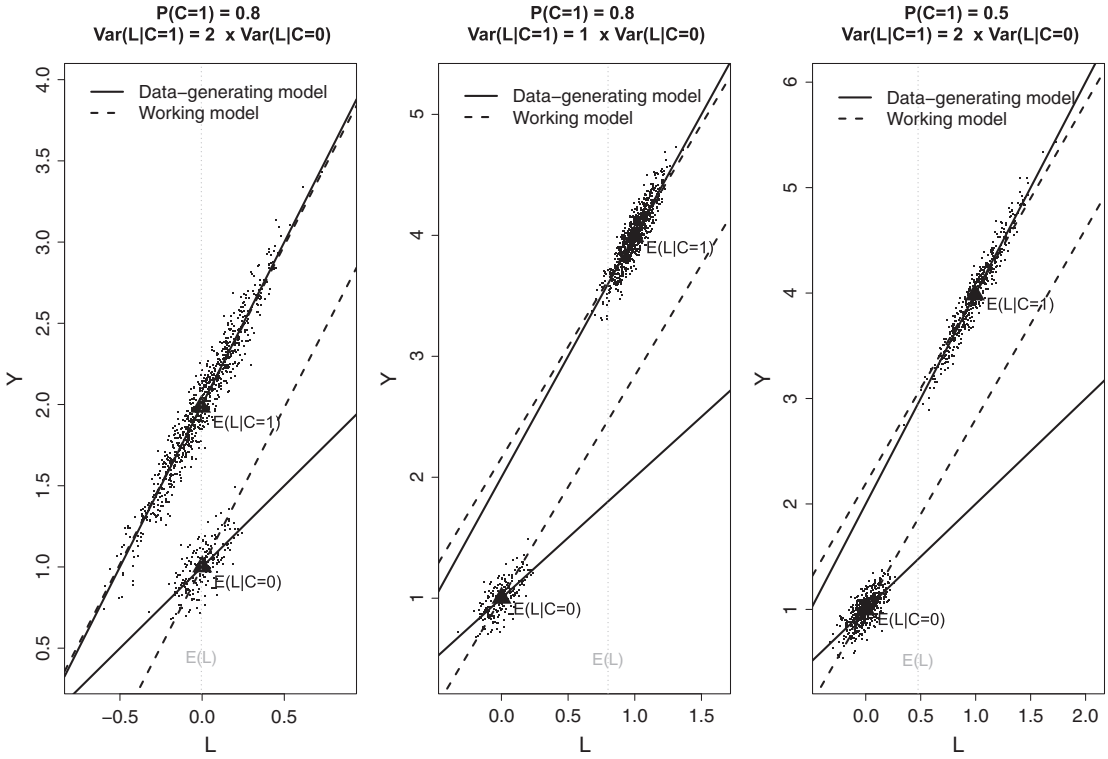


Figure 2. Regression line for the data-generating model (1) and for the working model (8) both with linear link function, considering two centers (center 0 at the bottom and center 1 on top) and a scalar L.

anyway, and the risk for these patients under the care level of the other center is only slightly biased. So, the expected risk (7) for this center also has a small bias as we average the expected risks for that center's patients over all observed care levels.

The impact of the variance of the center-specific L-distribution is best understood when looking at two centers in (11) and (12). First, the larger the variance in a center's patient mix is, the greater this center's influence on the estimates of the working model parameters. Second, when for a given center the patient mix has large variance, the performance of the other center will be extrapolated to more extreme values of L for which there may be no good fit (Figure 2, left or right panel). Then, due to different extrapolation, the bias on the directly standardized risk for a given center increases with smaller variance in its patient distribution compared with other centers, while for indirect standardization the opposite is true.

In general, the bias on the indirectly standardized risk will be smaller than on the directly standardized risk, and this difference will be more apparent when there are many centers. This is because for indirect standardization, we take an average over all centers in (10), canceling out positive and negative values, but not for direct standardization in (9).

In summary, when ignoring interactions between center and a patient covariate L, we only expect large bias in the presence of large interaction effects and large differences in the center-specific mean of L across centers for both standardization techniques. Moreover, for direct standardization, the bias is expected to be the largest for the smallest centers and centers with small variance of the center-specific L-distribution. For indirect standardization, the bias is expected to be the largest for the largest centers and centers with large variance of the center-specific L-distribution.

In the Supporting Information, we investigate whether bias can be reduced by using model-based estimators for $E(Y)$ and $E\{Y(c)|C = c\}$. We find that in comparisons with the directly standardized risk $E\{Y(c)\}$, it is not always beneficial to use the model-based overall mortality. Bias on the indirectly standardized risk will never be reduced by using the model-based estimator for $E\{Y(c)|C = c\}$.

4. Simulation study

We perform a simulation study to assess the impact of ignoring interactions in logistic regression models. Besides studying the bias, we also examine efficiency in terms of the mean squared error (MSE) on the standardized risk of interest.

We simulate $S = 500$ datasets with $n = 10\,000$ patients distributed over $m = 50$ centers and including a scalar patient-specific covariate L. We first generate the patient characteristic, for example, scaled age, following a standard normal $N(0, 1)$ or right-skewed Beta(1,6) distribution and assign each patient to a specific center c, following the propensity score model

$$P(C = c|L) = \frac{\exp\left(\alpha_{0c} + \alpha_{1c}L\right)}{\sum_{j=1}^{m} \exp\left(\alpha_{0j} + \alpha_{1j}L\right)}, \tag{13}$$

where α_{0c} determines the relative center size. Differences in patient mix are large when we impose strongly varying values of α_{1c} among centers. For this study population, we generate a binary outcome Y, for example, 30-day mortality, following a logistic outcome regression model as in (1) with the logistic link function. In the Supporting Information, we plot the mortality risk in function of L for each center to illustrate the magnitude of the interaction effect. There, we also describe the center-specific distribution of the marginally normal standardized or beta distributed covariate L with small or large differences across centers for one simulated dataset.

Bias and precision for the standardized risks are estimated for a working model, which includes interactions between L and C or not. Model parameters are estimated using Firth-corrected maximum likelihood methods. For convenience, we denote the directly or indirectly standardized risk for center c based on the data-generating model by f_c, based on the fitted working model in simulation run $s = 1, \ldots, S$ by $\hat{f}_c^s$, and the average of the latter over all simulation runs by $\hat{f}_c$. The center-specific bias on the directly or indirectly standardized risk for center c is then estimated by

$$S^{-1} \sum_{s=1}^{S} \left(\hat{f}_c^s - f_c\right) = \hat{f}_c - f_c. \tag{14}$$

To prevent that positive bias in some centers is canceled out by negative bias in other centers, we square these center-specific biases before taking the average over all centers. Then, the overall bias is defined as follows:

$$\sqrt{m^{-1}\sum_{c=1}^{m}\left(\hat{f}_c - f_c\right)^2 - S^{-1}\hat{\mathrm{Var}}\left(\hat{f}_c^s\right)}, \tag{15}$$

which includes a penalty because the average of the estimated squared bias is partly influenced by the imprecision of the estimates. Here, $\hat{\mathrm{Var}}\left(\hat{f}_c^s\right)$ denotes the estimated variance of the center-specific standardized risk over all simulation runs. The square root of the overall mean squared error is estimated by

$$\sqrt{m^{-1}\sum_{c=1}^{m}S^{-1}\sum_{s=1}^{S}\left(\hat{f}_c^s - f_c\right)^2}. \tag{16}$$

Finally, we measure the variability in patient mix across centers by the variance of the random intercepts in a random intercept model for L conditional on center.

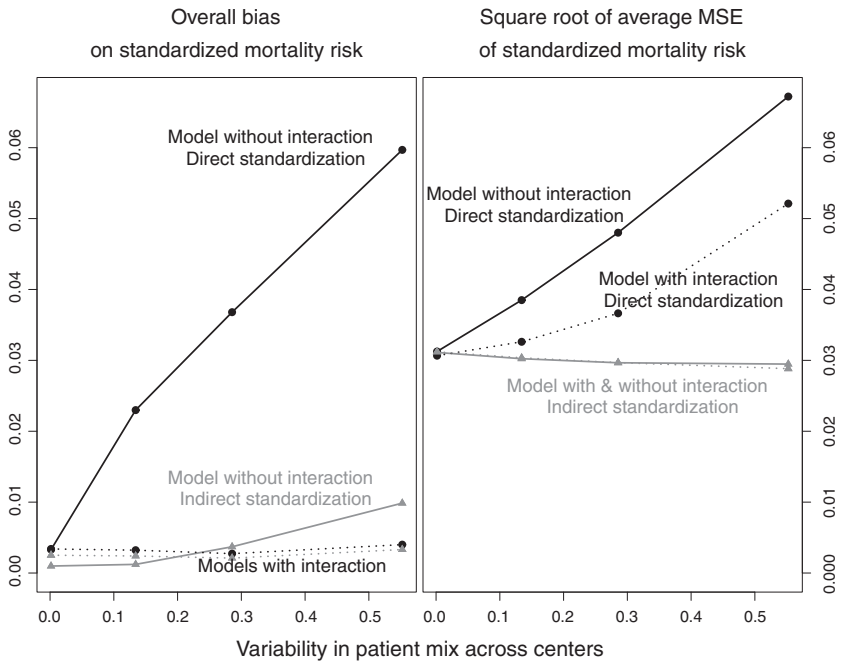
For normally distributed L, the simulation results are similar to the earlier theoretical findings in Section 3 (Figure 3(a)). That is, when there is little confounding by L, the models with and without interactions give comparable standardized risks. However, when patient mix differs much across centers, the overall bias and root mean squared error for the directly standardized risk $E\{Y(c)\}$ are large. For indirect standardization, the overall bias is not as large as for direct standardization, as explained earlier, and the MSE seems insensitive in excluding the interactions. Here, the variance has by far the largest contribution in the MSE, so apparently, precision on the indirectly standardized risk barely changes when ignoring the center-patient interactions. The overall bias following the model with interactions is sometimes larger than without interactions, which may be due to small sample bias or overfitting problems.

In the Supporting Information, we show results for one simulated dataset and indeed detect most bias when there are large differences in patient mix. For directly standardized risks, the smallest centers suffer most from bias when ignoring center-patient interactions, while for indirectly standardized risks, the largest centers may still suffer substantial bias. It can also be seen that the direction of this bias is not necessarily so that it shrinks the estimated outcome more toward the overall mean or zero, which would mask centers from being detected as having outlying performance. In practice, we thus do not know a priori whether the center's performance is overestimated or underestimated because of ignoring the interactions.

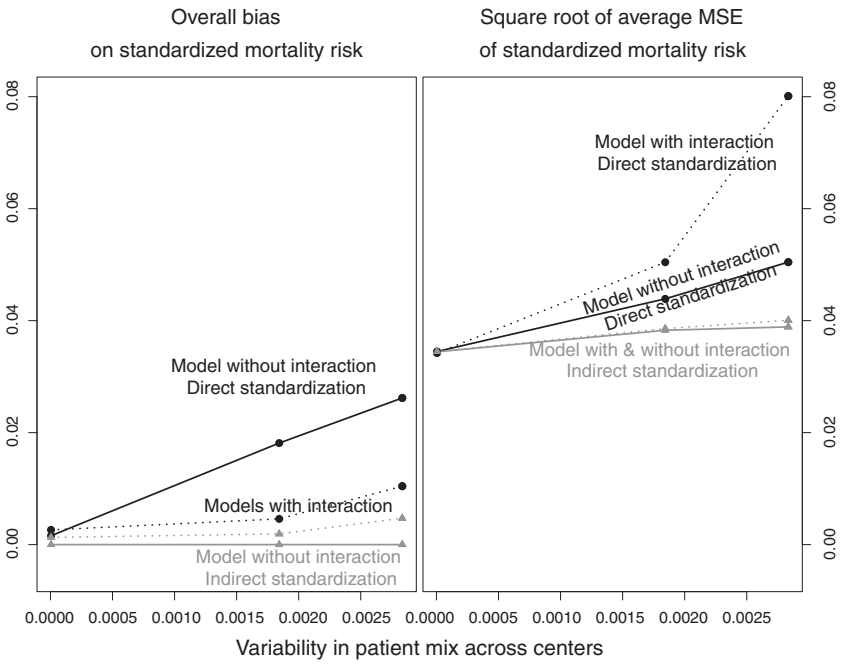
Surprisingly, for beta distributed L in Figure 3(b), we see less bias for the model without interactions than with interactions for indirect standardization. It is also remarkable that for direct standardization and large variability in patient mix, the MSE is larger for the model with interactions than for the model without interactions. Both these findings are due to fitting problems when modeling interactions with beta distributed L.

5. Data analysis: Riksstroke, the Swedish national quality register for stroke care

Riksstroke (http://www.riksstroke.org/eng) is a national quality register for acute stroke, collecting data from all 90 Swedish hospitals. The register contains 249 414 adult patients (⩾18 years) with first registered stroke between 2001 and 2012. We consider patients diagnosed with ischemic stroke (ICD-10 I63), intracerebral hemorrhage (ICD-10 I61), or unspecified acute cerebrovascular event (ICD-10 I64). Centers are compared in terms of directly or indirectly standardized 30-day mortality risks that correct for the patients' sex, age, level of consciousness at arrival (alert, drowsy, or unconscious), which is a proxy for baseline severity, and time to hospital (hours between stroke and arrival at hospital). The latter could be an important predictor because brain tissue is rapidly lost as stroke progresses, and the sooner treatment (e.g., thrombolysis) is initiated, the larger the probability of a favorable outcome [18]. The observed time to hospital was more than 24 h for several patients, which are thought to be mistakes in the registration and therefore truncated at 24 h. Interactions with patient's age may arise when some centers make special efforts for the revalidation of older patients. Differences in center performance may also differ across groups of time until hospital arrival depending on differences in prenotification systems [19] and time from arrival to thrombolysis treatment. Riksstroke typically reports directly standardized risks as



(a) Standard normal distribution for L.



(b) Beta distribution for L.

Figure 3. Estimated bias and precision for direct and indirect standardization are based on $S = 500$ simulations. Black dots are used for direct standardization and gray triangles for indirect, full lines are used for models without interactions and dotted lines for models with interactions. (a) Standard normal distribution for L and (b) Beta distribution for L.

one aims to compare intrinsic qualities of the centers. Here, we will estimate directly and indirectly standardized risks as we aim to provide insight in the bias in the setting studied here, where each of both standardization techniques could be of interest, depending on the research question posed.

There are 2 852 records with missing consciousness level and 148 910 with missing time to hospital. We discuss the results of two different approaches to handle these missing data: (1) We assume that the data are missing completely at random and perform a complete case analysis. To prevent quasi-complete separation, we also exclude the two smallest centers (center size 5 and 22) with respectively 0 and 1 death

within 30 days after admission. This resulted in a reduction of the dataset to 100 207 records, and overall 30-day mortality risk decreased from 13.13% to 12.48%. (2) We assume that the data are missing at random and perform five imputations of the missing data using the R-package Multivariate Imputation by Chained Equations (MICE) [20]. A description of the predictors that were used for the imputation models is given in the Supporting Information. As we need to allow for interactions with center in the outcome model, we fit separate imputation models per center, with center sizes ranging from 56 to 11 669 (Median 2324). No outcome values were missing.

Standardized outcomes were based on a Firth-corrected fixed effects model, with or without interactions between center and time to hospital or age. To suggest a functional form for time to hospital in the outcome regression model while accounting for the other prognostic factors, we categorized time following its 10% quantiles in the model without interactions. We found a good fit for a loglinear effect of time, and similarly for a linear and quadratic age effect. We found that for longer time to hospital, the 30-day mortality risk decreased for alert patients, while for drowsy or unconscious patients, the risk increased (Figure in Supporting Information). Therefore, we will allow for an interaction between time to hospital and consciousness level in all fitted outcome models. The decreased risk for alert patients may be due to different baseline severity within this group of patients: Patients with a less severe stroke have lower mortality risk but it also takes longer to recognize the symptoms and reach the hospital, while patients with obvious symptoms arrive earlier but also have a higher mortality risk. For drowsy or unconscious patients, the symptoms are more apparent, and patients who arrive early will have a lower mortality risk than those who arrive late.

In summary, we fit three models; the first includes the center where the patient was treated, patients' sex, age and quadratic age, level of consciousness and log transformed time to hospital as main effects, and an interaction between consciousness and time to hospital. The other two models additionally allow the center effect to depend on the linear effect of age or the loglinear effect of time to hospital.

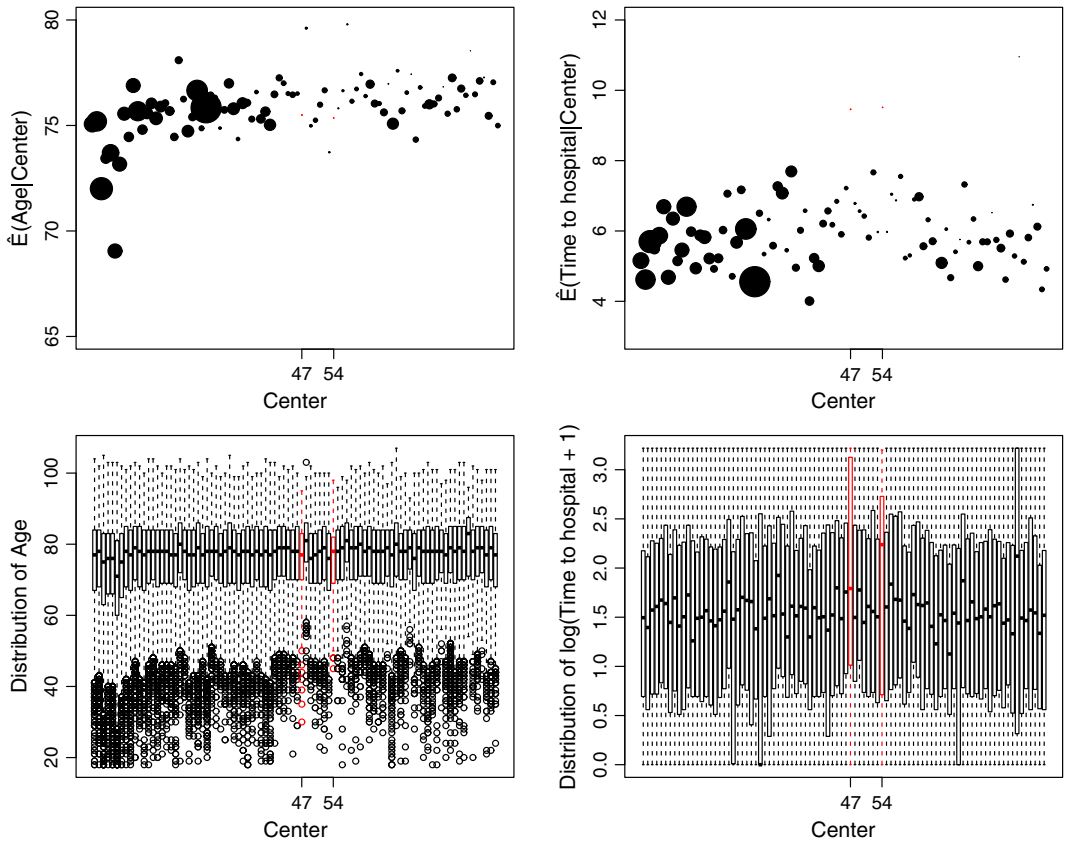


Figure 4. Center-specific values for the patient's age and time to hospital (hours) for one imputed dataset. Bubble size is proportional to center size. Center 47 and 54 have more than 1% difference in its estimated potential full population risk when ignoring interactions with time to hospital (MI).

An overall Wald test for interactions with center was obtained for age (p-value 0.009 for CC and < 0.001 for MI) and for time to hospital (p-value < 0.001 for complete cases (CC) and multiple imputed data (MI)). We will now investigate how the standardized risks differ when based on a model with or without interactions between center and patient's age or time to hospital. Results are based on the CC or the MI, where for the latter, we report results averaged over five imputed datasets unless otherwise stated.

We see substantial differences in patient mix across centers coming from time to hospital (Figure 4 for MI, Figure for CC in Supporting Information), and only minor differences are seen for age. We measure the variability in patient mix across centers as before, that is, by the variance on the random intercepts in a random intercept model for L conditional on center, and averaged over the imputed datasets, we obtain 0.022 (CC) or 0.024 (MI) for standardized log time to hospital compared with 0.017 (CC) or 0.013 (MI) for standardized age. So from previous theoretical findings, we know that the model ignoring interactions

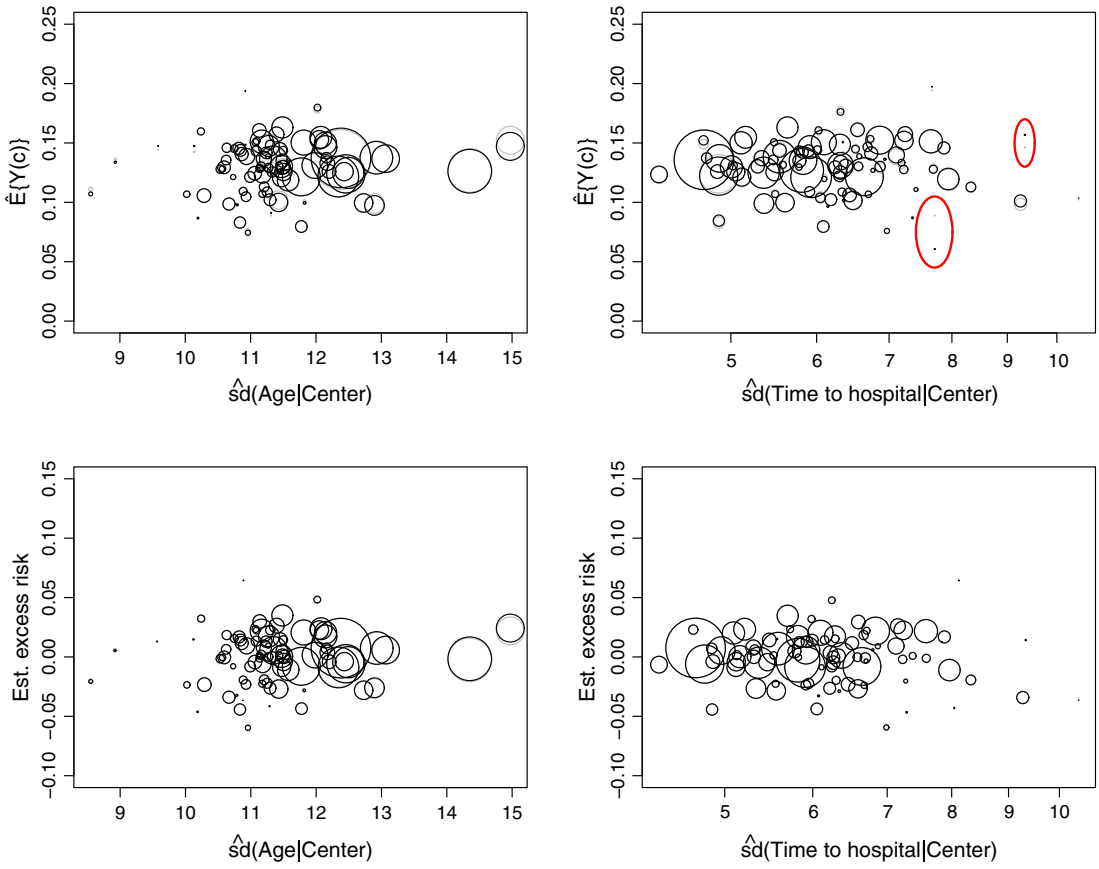


Figure 5. The directly or indirectly standardized risk per center, with or without interactions between center and patient's age or time to hospital (grey without and black with interactions), in function of the standard deviation of the center-specific distribution of patient's age or time to hospital for multiple imputation analysis. Bubble size is proportional to center size and ellipses indicate centers with more than 1% difference in estimated mortality risk.

Table I. The difference in estimated standardized risk between the model with and without interactions between center and patient's age or time to hospital, based on complete cases (CC) or multiple imputed data (MI). We report the maximum difference, the average difference (square root of the average of squared differences), and the number of centers for which the difference in standardized risk exceeds 1%.

		Max. difference (%)		Average difference (%)		No. centers with difference > 1%	
		CC	MI	CC	MI	CC	MI
Direct stand.	Age	1.31	0.63	0.26	0.14	1	0
	Time to hospital	1.33	2.83	0.28	0.35	2	2
Indirect stand.	Age	0.22	0.24	0.05	0.04	0	0
	Time to hospital	0.37	0.12	0.09	0.03	0	0

with time to hospital rather than age may induce larger bias on the standardized risks, although the bias will be minor for both (Figure 3(a)).

In general, we see negligible differences in standardized risk when based on a model without or with interactions between center and either age or time to hospital (Figure 5 for MI, Figure for CC in Supporting Information). However, for the direct standardization, we found two (CC) or two (MI) centers with a difference of more than 1% in risk when ignoring interactions with time to hospital and one (CC) center when ignoring interactions with age. As expected, these differences are larger for direct compared with indirect standardization (Table I). In addition, these differences are larger for time to hospital than for age. So, although for some centers we found a strong interaction with age, the standardized risks were found to be more robust because the age distribution does not differ much across centers.

6. Discussion

We found that if some centers actually perform better on a specific group of patients compared with other centers, then ignoring this in the analysis may bias the directly and indirectly standardized risks when the corresponding patient characteristic is very differently distributed between centers, but bias is negligible otherwise. We therefore advise special attention to interactions with covariates whose distribution differs substantially across centers. When there is no large variability in patient mix, then the common practice of ignoring center-patient interactions does not severely impact standardized mortality risks. In general, we notice larger bias for directly standardized compared with indirectly standardized risks. However, for directly standardized risks, the largest bias is seen for centers with the smallest proportion of registered patients as opposed to the larger centers for indirectly standardized risks. In our study, we found the same trends for the overall root mean squared error. Of course the interaction effect will need to claim its role when interest lies in prediction of the mortality risk for a specific subgroup rather than directly or indirectly standardized risks.

To detect centers with low or high mortality risks, we have applied a similar center classification technique as in [7] on the simulated data (Supporting Information). A center is classified as low/high risk if the data provide sufficient evidence that the standardized risk exceeds a clinical benchmark, for example, relative to the population average risk $E(Y)$. We found that ignoring center-patient interactions has similar impact on correct center classification as on the bias: the largest differences are seen for direct standardization and large differences in case mix across centers (Supporting Information). Surprisingly, the power to detect outlying performance is not always decreased by mistakenly ignoring the interactions, but then we see more centers wrongly classified as having outlying performance. Reassuringly, in general, the percentage of correct center classification is very similar for the model with and without interactions and this is both for direct and indirect standardization.

We expect that the impact of center-patient interactions on the standardized risk depends on the considered disease. For example, for a register on a non-acute surgical procedure, we may expect a large impact. First, patient mix may differ substantially across hospitals when patients can choose the hospital where they are treated. Moreover, treatment and thus the patient's mortality risk is partly based on the surgeons' decisions and experience, which makes it more subject to effect modification, for example, when some hospitals are less experienced with a specific surgery. On the other hand, for a register on acute stroke, patients are mostly treated at the nearest hospital so there is less confounding. Furthermore, well-defined treatment guidelines for this disease result in the same procedure given in each clinical center, thus we expect the difference in mortality risk between, for example, old and young patients to be similar across centers.

In practice, it is not always possible to estimate all interaction parameters, especially when the number of patient characteristics is large. One option is to use prior knowledge on hospital specialization and reduce the factors for which interactions may be considered. In addition or alternatively a summary measure for the patient's baseline severity may reduce the number of parameters to be estimated, for example, in the form of propensity scores or prognostic scores [21, 22]. We used penalized likelihood estimation, more specifically the Firth correction to overcome fitting problems. In this context random effects models are often used which help to reduce the effective model dimension when allowing for differential effects of patient characteristics across centers [1, 10]. However, it has been repeatedly shown that the power for detecting outlying center performance is much lower when using normal random center effects compared with Firth-corrected fixed effects [7, 23]. In future work, it may be of interest to investigate whether more general regularization methods bring a solution.

Acknowledgement

This work was supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) [to M.V.]; IAP research network from the Belgian government (Belgian Science Policy) [grant no. P07/06 to E.G. and S.V.], and the Swedish Research Council [grant no. 2012-5934 to E.G. and M.E.]. *Conflict of Interest*: None declared.

References

1. Normand S, Glickman M, Gatsonis C. Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association* 1997; **92**(439):803–814.
2. Shahian DM, Blackstone EH, Edwards FH, Grover FL, Grunkemeier GL, Naftel DC, Nashef SA, Nugent WC, Peterson ED. Cardiac surgery risk models: a position article. *The Annals of Thoracic Surgery* 2004; **78**(5):1868–1877.
3. Shahian D, Normand S. Comparison of “risk-adjusted" hospital outcomes. *Circulation: Journal of the American Heart Association* 2008; **117**:1955–1963.
4. Spiegelhalter D. Funnel plots for comparing institutional performance. *Statistics in Medicine* 2005; **24**(8):1185–1202.
5. DeLong E, Peterson E, DeLong D, Muhlbaier L, Hackett S, Mark D. Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine* 1997; **16**(23):2645–2664.
6. He K, Kalbfleisch JD, Li Y, Li Y. Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. *Lifetime Data Analysis* 2013; **19**(4):490–512.
7. Varewyck M, Goetghebeur E, Eriksson M, Vansteelandt S. On shrinkage and model extrapolation in the evaluation of clinical center performance. *Biostatistics* 2014; **15**(4):651–664.
8. Gatsonis C, Normand SL, Liu C, Morris C. Geographic variation of procedure utilization: a hierarchical model approach. *Medical Care* 1993; **31**(5):YS54–YS59.
9. Gatsonis CA, Epstein AM, Newhouse JP, Normand SL, McNeil BJ. Variations in the utilization of coronary angiography for elderly patients with an acute myocardial infarction: an analysis using hierarchical logistic regression. *Medical Care* 1995; **33**(6):625–642.
10. Austin P, Alter D, Tu J. The use of fixed- and random-effects models for classifying hospitals as mortality outliers: a Monte Carlo assessment. *Medical Decision Making* 2003; **23**(6):526–539.
11. Saposnik G, Baibergenova A, O’Donnell M, Hill M, Kapral M, Hachinski V and On behalf of the Stroke Outcome Research Canada (SORCan) Working Group. Hospital volume and stroke outcome does it matter? *Neurology* 2007; **69**(11): 1142–1151.
12. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**(1):29–46.
13. VanderWeele TJ, Mukherjee B, Chen J. Sensitivity analysis for interactions under unmeasured confounding. *Statistics in Medicine* 2012; **31**(22):2552–2564.
14. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health* 2006; **60**(7):578–586.
15. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**(1):27–38.
16. Shahian DM, Normand SL, Torchiana DF, Lewis SM, Pastore JO, Kuntz RE, Dreyer PI. Cardiac surgery report cards: comprehensive review and statistical critique. *The Annals of Thoracic Surgery* 2001; **72**(6):2155–2168.
17. Liu J, Gustafson P. On average predictive comparisons and interactions. *International Statistical Review* 2008; **76**(3): 419–432.
18. The ATLANTIS, ECASS, and NINDS rt-PA Study Group Investigators. Association of outcome with early stroke treatment: pooled analysis of atlantis, ecass, and ninds rt-pa stroke trials. *Lancet* 2004; **363**(9411):768–774.
19. Lin CB, Peterson ED, Smith EE, Saver JL, Liang L, Xian Y, Olson DM, Shah BR, Hernandez AF, Schwamm LH, Fonarow GC. Emergency medical service hospital prenotification is associated with improved evaluation and treatment of acute ischemic stroke. *Circulation: Cardiovascular Quality and Outcomes* 2012; **5**(4):514–522.
20. Buuren S, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. *Journal of Statistical Software* 2011; **45**(3):1–67.
21. Rubin D. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 1997; **127**(8S):757–763.
22. Hansen BB. The prognostic analogue of the propensity score. *Biometrika* 2008; **95**(2):481–488.
23. Kalbfleisch JD, Wolfe RA. On monitoring outcomes of medical providers. *Statistics in Biosciences* 2013; **5**(2):286–302.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher’s web site.

