Adjusting for Disease Severity Across ICUs in Multicenter Studies

Timo B. Brakenhoff, PhD[1]; Nienke L. Plantinga, MD, PhD[1]; Bastiaan H. J. Wittekamp, MD, PhD[1,2]; Olaf Cremer, MD, PhD[2]; Dylan W. de Lange, MD, PhD[2,3]; Nicolet F. de Keizer, PhD[3,4]; Ferishta Bakhshi-Raiez, PhD[3,4]; Rolf H. H. Groenwold, MD, PhD[1,5]; Linda M. Peelen, PhD[1]

Objectives: To compare methods to adjust for confounding by disease severity during multicenter intervention studies in ICU, when different disease severity measures are collected across centers.

Design: In silico simulation study using national registry data.

Setting: Twenty mixed ICUs in The Netherlands.

Subjects: Fifty-five–thousand six-hundred fifty-five ICU admissions between January 1, 2011, and January 1, 2016.

Interventions: None.

Measurements and Main Results: To mimic an intervention study with confounding, a fictitious treatment variable was simulated whose effect on the outcome was confounded by Acute Physiology and Chronic Health Evaluation IV predicted mortality (a common measure for disease severity). Diverse, realistic scenarios were investigated where the availability of disease severity measures (i.e., Acute Physiology and Chronic Health Evaluation IV, Acute Physiology and Chronic Health Evaluation II, and Simplified Acute Physiology Score II scores) varied across centers. For each scenario, eight different methods to adjust for confounding were used to obtain an estimate of the (fictitious) treatment effect. These were compared in terms of relative (%) and absolute (odds ratio) bias to a reference scenario where the treatment effect was estimated following correction for the Acute Physiology and Chronic Health Evaluation IV scores from all centers. Complete neglect of differences in disease severity measures across centers resulted in bias ranging from 10.2% to 173.6% across scenarios, and no commonly used methodology–such as two-stage modeling or score standardization–was able to effectively eliminate bias. In scenarios where some of the included centers had (only) Acute Physiology and Chronic Health Evaluation II or Simplified Acute Physiology Score II available (and not Acute Physiology and Chronic Health Evaluation IV), either restriction of the analysis to Acute Physiology and Chronic Health Evaluation IV centers alone or multiple imputation of Acute Physiology and Chronic Health Evaluation IV scores resulted in the least amount of relative bias (0.0% and 5.1% for Acute Physiology and Chronic Health Evaluation II, respectively, and 0.0% and 4.6% for Simplified Acute Physiology Score II, respectively). In scenarios where some centers used Acute Physiology and Chronic Health Evaluation II, regression calibration yielded low relative bias too (relative bias, 12.4%); this was not true if these same centers only had Simplified Acute Physiology Score II available (relative bias, 54.8%).

Conclusions: When different disease severity measures are available across centers, the performance of various methods to control for confounding by disease severity may show important differences. When planning multicenter studies, researchers should make contingency plans to limit the use of or properly incorporate different disease measures across centers in the statistical analysis. (*Crit Care Med* 2019; 47:e662–e668)

Key Words: Acute Physiology and Chronic Health Evaluation; computer simulation; confounding factors; disease severity; intensive care units; Simplified Acute Physiology Score

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands.

[2] Department of Intensive Care Medicine, University Utrecht, The Netherlands.

[3] National Intensive Care Evaluation (NICE) Foundation, Amsterdam, The Netherlands.

[4] Department of Medical Informatics, Amsterdam UMC Amsterdam Public Health Research Institute, University of Amsterdam, Amsterdam, The Netherlands.

[5] Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands.

Dr. Brakenhoff and Dr. Plantinga contributed equally.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).

Dr. de Keizer's institution received funding from National Intensive Care Evaluation foundation, during the conduct of the study (grants and advisory board member). Dr. Groenwold was supported financially by the Netherlands Organisation for Scientific Research (project 917.16.430). The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: N.L.Plantinga@umcutrecht.nl

DOI: 10.1097/CCM.0000000000003822

The prognosis of ICU patients is influenced by disease severity at the time of ICU-admission. Several prediction models have been developed to quantify disease severity and predict hospital survival, among ICU patients. Examples include the Acute Physiology and Chronic Health Evaluation (APACHE II, III, and IV) scores and the Simplified Acute Physiology Score (SAPS II and III) (1–4). When analyzing the effects of interventions in the ICU, these measures are often used as a proxy for actual disease severity to correct for potential confounding.

However, different ICUs may routinely collect different disease severity measures (DSMs). When there is not a single, common score measured across centers, adjustment for confounding may not be straightforward. This applies to observational multicenter studies, cluster-randomized trials, and individual participant data meta-analyses (5–7). Although numerous studies have evaluated the prognostic performance of different DSMs across various settings (8–10), the aim of the current study was to compare different methods to adjust for confounding by disease severity in multicenter ICU studies when different measures are available from different centers, and to assess how the performance of these methods depends on the availability of these measures across centers.

METHODS

Study Design

A simulation study was performed using data from the Dutch "National Intensive Care Evaluation (NICE)" registry (11). This database holds information on APACHE II, APACHE IV, and SAPS II scores and their predicted mortalities as well as observed in-hospital mortality status for all admissions in all ICUs in the Netherlands (see *Data* section). To mimic an observational study with confounding, where sicker patients are more likely to receive the treatment under study, a fictitious treatment was assigned to half of the patients in each center. Treatment was assigned conditional only on the APACHE IV predicted mortality. In doing so, treatment status is independent of the outcome when correcting for the APACHE IV predicted mortality, and therefore the corrected odds ratio (OR) of treatment on in-hospital mortality is 1.

Thereby, the treatment was noneffective and any deviation from the OR of 1.0 was caused by bias due to incomplete confounding adjustment (see *Simulating Treatment* section). Subsequently, multiple scenarios were evaluated, in which the hypothetical availability of DSMs differed across centers depending on patient- and center-level characteristics (see *Scenarios* section). Within each scenario, several methods to adjust for confounding (see *Confounding Adjustment Methods* section) were applied to estimate the association of the simulated treatment with the outcome, adjusted for the available DSMs. Confounding adjustment by APACHE IV score in all centers was chosen as the reference method (scenario 0) to which other combinations of methods and scenarios were compared, because we assume that these scores are much more frequently available to clinical studies than the APACHE IV predicted mortalities (these predictions derive from complex calculations based on many separate variables) (12, 13). Thereby, we aim to identify "best-case" methods, which are also applicable to practice. To account for simulation error, each scenario was repeated 1,000 times on a different bootstrap sample from the original data, results were averaged and methods compared in terms of bias in the estimated treatment effect and the coverage of its 95% CI (see *Performance Measures section*). Simulations were performed using R (Version 3.2.2; R Foundation for Statistical Computing, Vienna, Austria; https://www.R-project.org/). The NICE registry is registered according to the General Data Protection Regulation. The medical ethics committee of the Amsterdam UMC stated that medical ethics approval for this study was not required under Dutch national law (registration number W18_179).

Data

The NICE registry contains information of more than 80,000 ICU admissions per year from all 84 Dutch ICU centers (11, 14). To obtain a generalizable selection of hospitals with differences in level of care, volume, and case-mix, 10 university/teaching centers ("teaching") and 10 peripheral ("nonteaching") centers were chosen randomly. Herein, all unique hospital admissions admitted to the ICU between January 1, 2011, and January 1, 2016, were extracted. Subsequently, patients younger than 18 years, admissions with ICU length of stay less than 24 hours, planned admissions for chronic respiratory disease, and records with missing information on disease severity were excluded ($n = 318$). These criteria were chosen to mimic a possible cohort of a medical intervention study in the ICU, and therefore do not correspond to the patient selection criteria for which the models were developed.

Simulating Treatment

A fictitious treatment (predetermined OR, 1.0) was assigned to half of the patients per center, based on each individual's APACHE IV predicted mortality. The APACHE IV model was chosen, as it is the most recent prediction model with most diagnostic categories. Furthermore, it is most inclusive with regard to specific populations (i.e., cardiac surgery patients) (1, 15, 16). The data-generating model (DGM) used to relate APACHE IV predicted mortality to the probability of treatment is:

$$\text{logit}[P(X_i = 1)] = \alpha + \beta_{AP4} AP4i$$

where $AP4_i$ stands for APACHE IV predicted mortality of individual i and X_i denotes treatment status of individual i (0 = untreated, 1 = treated). To ensure that half of the patients in the ICU received treatment, a search procedure was applied using the original data to determine what the value of the intercept α in equation 1 should be; –0.28. Coefficient β_{AP4} was set to 1.25, corresponding to a crude (i.e., not adjusted for confounding) OR of 1.50 for the effect of the treatment on in-hospital mortality. Such a confounding effect has been observed in earlier publications (17). This effect was determined using an iterative procedure using the original data, where β_{AP4} was incrementally changed to achieve the desired crude effect. By sampling from a Bernoulli distribution with probability of success P (*X*i = 1), a treatment status was assigned to each individual, independent of in-hospital mortality and conditional on the APACHE IV predicted mortality.

Scenarios

In our reference scenario—to which all subsequent scenarios were compared—all centers were assumed to have the APACHE IV score available (scenario 0).

Nine scenarios were investigated, which differed with regards to the availability of the APACHE IV score or the alternative score across centers, based on center and case-mix characteristics (**Table 1**; and **Supplementary Table 1**, Supplemental Digital Content 1, http://links.lww.com/CCM/E603). These characteristics include the type of center (teaching vs nonteaching), center volume (> 1,500 vs < 1,500 included admissions in 2011–2015), average APACHE IV predicted probability of in-hospital death (> 25% vs < 25%), and proportion of medical admissions (> 70% vs < 70%).

To investigate confounding adjustment using both the APACHE II score and the SAPS II score relative to the APACHE IV score, two separate analyses were performed for all nine scenarios: participating centers without the APACHE IV score had the APACHE II score (A) or the SAPS II score (B) available (Table 1).

Confounding Adjustment Methods

Eight different methods to estimate the effect of treatment on in-hospital mortality were considered, each unique in their strategy to incorporate two distinct DSMs for confounding adjustment (**Table 2**). Each method had a common fixed effects logistic regression as the analysis model, which included as independent variables the assigned treatment status (X), a measure of disease severity (as determined by the different methods), and a fixed intercept per center. A short description of each method is given in Table 2.

Performance Measures

The procedure described above was executed on each of 1,000 bootstrap samples from the selected data, where patients were bootstrapped within centers. Within each method and each scenario, the estimated treatment effects were averaged and compared with the treatment effects found in scenario 0 (reference effect: adjustment with APACHE IV score in all centers), which was assumed to be without uncertainty. For each of methods one through eight with scenario's one through nine, we calculated mean effect estimates over the bootstrap estimates and reported 95% percentile CIs. Subsequently, bias was calculated for each of these as follows: relative bias = (mean effect estimate–reference effect)/reference effect, absolute bias = mean effect estimate–reference effect (the absolute bias was reported on the OR scale by exponentiating this formula). Coverage was calculated by observing the proportion of times the reference effect (treatment effect of scenario 0) fell within the 95% CI constructed around the effect estimate obtained for each bootstrap sample.

RESULTS

The resulting dataset consisted of 55,655 ICU admissions (**Table 3**). The median number of average yearly admissions

TABLE 1. Description of the Availability of Disease Severity Measures Across Centers for Each Scenario

Scenario	Availability of Disease Severity Measures Across Centers	
0 (reference)	All centers had the APACHE IV score	
	APACHE IV score	**Analysis A: APACHE II score** **Analysis B: SAPS II score**
1	Teaching centers	Nonteaching centers
2	Nonteaching centers	Teaching centers
3	High-volume centers[a]	Low-volume centers
4	Low-volume centers	High-volume centers[a]
5	APACHE IV highest risk[b]	APACHE IV lowest risk
6	APACHE IV lowest risk	APACHE IV highest risk[b]
7	High proportion of medical admissions[c]	Low proportion of medical admissions
8	Low proportion of medical admissions	High proportion of medical admissions[c]
9	APACHE IV and APACHE II/SAPS II distributed approximately evenly over centers[d]	

APACHE = Acute Physiology and Chronic Health Evaluation, SAPS = Simplified Acute Physiology Score.

[a] Centers with > 1,500 admissions in 2011–2015.

[b] Centers with mean APACHE IV predicted mortality > 25%.

[c] Centers with proportion medical admissions > 70%.

[d] Twelve centers had the APACHE IV score available and eight either the APACHE II or SAPS II score.

TABLE 2. Methods to Adjust for Confounding by Disease Severity, When Different Measures Are Available From Different Centers

Method	Description
1) Naive	Differences in DSMs across centers were neglected. As such, the disease severity scores were treated as if they had been measured in an identical fashion across centers and were combined into one disease severity variable included in the analysis model (i.e., falsely assuming that an APACHE IV of 30 corresponds to an APACHE II of 30).
2) Restriction	Only data from those centers that had the APACHE IV score available were included in the analysis.
3) Two-stage analysis	Each group of centers that had the same DSM available was analyzed separately. Subsequently, the estimated treatment effects and SEs were pooled by inverse variance weighting (18).
4) Cluster standardization	Within each group of centers with the same DSM, scores were standardized by setting the mean value of the empirical distribution to 0 with a SD of 1. The standardized disease severity scores were then combined into one variable and included as a covariate in the analysis model.
5) Center standardization	Within each center, DSMs were standardized by setting the mean value of the empirical distribution to 0 with a SD of 1. The standardized DSMs were then combined into one variable and included as a covariate in the analysis model.
6) Multiple imputation	When unavailable, APACHE IV scores were imputed, using information on treatment status, observed outcome, and the available (alternative) disease severity scores (i.e., APACHE II or SAPS II, which was assumed available in all centers). According to recent literature, including the outcome in the imputation model is considered appropriate and leads to more accurate imputations (19, 20). Each dataset was imputed five times using Bayesian linear regression (21, 22). Effect estimates and their SE were pooled using Rubin's rules (23).
7) Regression calibration	Again, APACHE II or SAPS II was assumed available in all centers. In the subset of centers that also had APACHE IV scores, the APACHE IV score was regressed on the treatment and the DSM available for all centers (APACHE II/SAPS II) using ordinary least-squares regression. The estimated coefficients of this regression model were then used to calibrate the estimated treatment effect and its SE, by means of the procedure described by Rosner et al (24).
8) Propensity score	Within each group of centers that had the same DSM available, a propensity score model—to predict the probability of having received treatment—was estimated. The estimated propensity scores were used to obtain inverse probability weights. A single weighted regression was then performed to estimate the treatment effect.

APACHE = Acute Physiology and Chronic Health Evaluation, DSM = disease severity measure, SAPS = Simplified Acute Physiology Score.

was 437 (interquartile range [IQR], 267–282) in all centers, 696 (IQR, 555–1,196) in teaching centers, and 238 (IQR, 174–325) in nonteaching centers. The average APACHE IV score was 68.5, 60.7, and 66.8, and the average APACHE IV predicted in-hospital mortality was 0.229, 0.236, and 0.203 for all, teaching, and nonteaching centers, respectively. There were six centers with more than 70% medical admissions, including three teaching and three nonteaching centers.

In **Supplementary Tables 2** and **3** (Supplemental Digital Content 1, http://links.lww.com/CCM/E603), we present the main results, being the performance of the different methods and scenarios in comparison with the reference scenario (scenario 0, adjustment for APACHE IV score); this is both presented as the relative bias (percentages) and as the absolute bias (on the OR scale). Negative percentages reflect ORs less than 1 and thereby bias toward a protective effect of treatment, whereas positive percentages reflect ORs greater than 1 and thereby bias toward a harmful treatment effect. Coverages are presented in **Supplementary Table 4** (Supplemental Digital Content 1, http://links.lww.com/CCM/E603).

In analysis A, where APACHE II scores were the alternative to APACHE IV scores, restriction, multiple imputation, and regression calibration resulted in the least bias, with a mean relative bias over the scenarios of 0.0%, 5.1%, and –12.4%, respectively (Supplementary Table 2, Supplemental Digital Content 1, http://links.lww.com/CCM/E603) and coverages above 92% for all methods (Supplementary Table 4, Supplemental Digital Content 1, http://links.lww.com/CCM/E603). The naive method gave, as expected, most relative bias with percentages ranging from 26.6% to 173.6% depending on the distribution of APACHE IV and APACHE II scores over centers, but with coverages less than 15% for six of eight scenarios. Standardization (both approaches), two-stage analysis, and propensity score modeling all resulted in comparable amounts of relative bias, with mean relative bias over the scenarios ranging from 41.5% to 48.6%.

In analysis B, where SAPS II scores were the alternative to APACHE IV scores, the least bias occurred with restriction and multiple imputation, with mean bias over the scenarios of 0.0% and 4.6% (Supplementary Table 2, Supplemental Digital Content 1, http://links.lww.com/CCM/E603) and coverages of 94.3% and 68.5%, respectively (Supplementary Table 4,

TABLE 3. Characteristics of the Selected Centers and Patients, Stratified for Teaching and Nonteaching Centers

Characteristics	Teaching Centers	Nonteaching Centers	All Centers
Number of included admissions (total)	43,516	12,139	55,655
Center characteristics[a]			
Number of centers	10	10	20
Number of high-volume centers (> 1,500 admissions in 2011–2015)	10	4	14
Number of centers with high average APACHE IV predicted mortality (> 25%)	6	0	6
Number of centers with high proportion of medical admissions (> 70%)	3	3	6
Average "yearly"[b] number of included admissions per center, median (IQR)	696 (555–1,196)	238 (174–325)	437 (267–682)
Patient characteristics			
Median age (IQR)	66 (55–74)	70 (59–78)	67 (55–75)
Male sex (%)	62	55	60
APACHE IV score, mean (SD)	68.5 (29.2)	60.7 (27.7)	66.8 (29.1)
APACHE IV predicted in-hospital mortality, mean (SD)	0.236 (0.255)	0.203 (0.227)	0.229 (0.250)
APACHE II score, mean (SD)	18.9 (7.67)	16.8 (7.75)	18.5 (7.74)
APACHE II predicted in-hospital mortality, mean (SD)	0.288 (0.248)	0.274 (0.228)	0.285 (0.244)
SAPS II score, mean (SD)	40.7 (16.9)	35.8 (16.9)	39.7 (17.0)
SAPS II predicted in-hospital mortality, mean (SD)	0.309 (0.267)	0.246 (0.249)	0.295 (0.265)
Died in hospital (%)	16.2	14.5	15.9

APACHE = Acute Physiology and Chronic Health Evaluation, IQR = interquartile range, SAPS = Simplified Acute Physiology Score.

[a] To preserve anonymity, information could not be presented for each individual center.

[b] From the ten centers, from which data from 5 years of admissions were used, indicates the average number of admissions per year per center. From these 10 averages are presented the median and IQR.

Supplemental Digital Content 1, http://links.lww.com/CCM/E603). The next best methods were standardization (both methods) and two-stage analysis, with relative bias of 18.4% and 18.7%, respectively—which are more than twice as low as in the scenarios with APACHE II—and high coverages. Interestingly, the propensity score method performed similar to the naive method, with relative biases of 33.2% and 36.2% and coverages of 83.6% and 81.4%, respectively. Contrary to the APACHE II scenarios, regression calibration resulted in high relative bias when some of the centers had SAPS II score, overestimating the treatment effect on average with 54.8% (coverage 68.5%). With regard to the scenarios, those where most centers or the teaching centers had APACHE IV resulted in least relative bias.

The ORs for all DSMs—when each is assumed available in all centers—are provided in **Supplementary Table 5** (Supplemental Digital Content 1, http://links.lww.com/CCM/E603). In comparison with the DGM, where treatment was simulated based on the APACHE IV predicted mortalities, the estimated treatment effect of scenario 0 was OR 1.08; this is the reference effect used to calculate bias for the different scenarios and methods, and it reflects the bias due to confounding adjustment by APACHE IV score rather than the APACHE IV predicted mortalities (OR, 1.00). For completeness, the obtained ORs and percentile CIs obtained for each method and scenario are presented in **Supplementary Table 6** (Supplemental Digital Content 1, http://links.lww.com/CCM/E603). Note that the absolute bias (on the OR scale) with respect to the DGM (correction using the APACHE IV predicted mortality; OR, 1.00) is easily obtained by subtracting 1 from all cells in this table.

DISCUSSION

In this simulation study, we compared eight methods to adjust for confounding by disease severity in multicenter ICU studies of medical interventions, where different DSMs are available across centers. Neglecting differences between DSMs across centers led to large relative bias in treatment effects. Commonly used methods such as two-stage modeling or standardization were unable to eliminate bias. Restriction of the analysis to centers where APACHE IV scores were available and multiple imputation of APACHE IV score resulted in the least bias. Regression calibration yielded low relative bias when some centers had APACHE II score available, but not when these had only SAPS II available.

The current analyses are by no means exhaustive and several assumptions were made to perform a simulation that would, in our opinion, best mimic a real-life multicenter study. First, the best proxy for disease severity was considered to be the APACHE IV predicted in-hospital mortality, which was used in the DGM to simulate treatment. We also assumed that in real-life studies, the best available DSM would be the APACHE IV score. This measure was not optimal, because correction for APACHE IV score in all centers (scenario 0) resulted in bias (OR, 1.08) relative to the DGM (Supplementary Table 5, Supplemental Digital Content 1, http://links.lww.com/CCM/E603). This may be explained by the fact that APACHE IV score does not have a linear relationship with in-hospital mortality; rather each variable and diagnostic category included in the model has a different relationship with the outcome, which is often ignored in confounding adjustment. Second, we assumed that there was a single average treatment effect (adjusted OR, 1.0) across centers, both in the DGM and in the analyses, and hence that there was no treatment effect heterogeneity across centers. In practice, however, there may be variation in treatment effects across centers. Inclusion of such variation may influence the results, but would likely not alter the conclusion. Instead, a different treatment effect variable—for example, an adjusted OR 1.5 (effective treatment) instead of OR 1.0—could influence the performance of the different methods (the amount of bias) in the different scenarios to a varying extent. Similarly, we assumed a single confounding effect across centers, resulting in a crude OR of 1.50; the magnitude of confounding may in practice vary with the type of study performed and may also differ across centers. Future studies may incorporate heterogeneity of treatment effects or confounding effects across centers by simulating the treatment using hierarchical regression models. Otherwise, data from a large multicenter ICU trial can also be used as the treatment variable is then already observed and does not have to be simulated.

This study has some limitations. The scenarios chosen in this simulation study partly overlapped; teaching hospitals, for example, were all high-volume centers (Supplementary Table 1, Supplemental Digital Content 1, http://links.lww.com/CCM/E603). This might explain some of the similarity of results across scenarios. Second, treatment was simulated based on APACHE IV predicted mortality, which was therefore by definition the only confounder in this study, while in practice, there may be more (unmeasured) confounding variables. Although the APACHE IV includes many different confounders such as age and comorbid conditions, it is uncertain to what extent our comparison of confounding adjustment methods is affected by that. Third, although our simulation study was based on a dataset with national coverage, results may not directly generalize to other countries with different case-mix (i.e., different distributions of disease severity across different countries). Fourth, with all centers trained to collect the various DSMs, we assumed no measurement error and no missing data; future research could assess the impact of these sources of bias on the different methods to control for confounding by disease severity in multicenter ICU studies. Finally, generalizability of our findings depends on the prognostic performance of the APACHE IV score in the respective setting, the underlying treatment effect size, and the type of outcome measure. The current study could be repeated with a different DSM to simulate treatment (e.g., recalibrated SAPS III in-hospital mortality probabilities), and with a yet other methods to correct for confounding. Furthermore, it should be repeated for different outcome types (continuous, time-dependent, instead of dichotomous) and with different magnitudes of treatment effectiveness.

In practice, when it is not feasible to collect a single DSM in all study centers, the choice for a specific confounding adjustment method may depend on the availability and type of DSM across centers. For multiple imputation and regression calibration, it is paramount that all centers have the alternative measure available. When that is not the case, the more modestly performing methods of standardization, two-stage analysis, and propensity score modeling may be best applicable. Although restriction resulted in least relative bias, it also yielded relatively imprecise estimates, which in practice could result in underpowered studies (which is unethical). Furthermore, in practice, studies may not want to restrict to a selection of sites based on the available DSM. As expected, the naive method, neglecting differences between APACHE IV and alternative scores, resulted in high relative bias, especially when APACHE IV score (range, 0–286) and APACHE II score (range, 0–72) were available (SAPS II score's range is 0–163). In practice, when choosing between different methods to control for heterogeneously measured confounders for the analysis of a particular dataset, this choice could be supported by small focused simulation studies for which our simulation code could be used as a starting point. The ultimate approach would be to try and obtain a single DSM from all centers, preferably the one that is most informative for the outcome of interest and treatment decisions in practice.

To the best of our knowledge, this is the first study to assess different modeling methods for confounding adjustment when different DSMs are available from different centers. Our results may help investigators of multicenter studies and individual participant data meta-analysis in the field of Intensive Care Medicine to design the statistical analysis when confounding adjustment is needed. Future research could compare these same methods in different datasets that may be smaller, have more measurement error, have heterogeneity in treatment effects, or have different DSMs available. In addition, investigators may focus on the performance of combinations of methods to most effectively adjust for confounding.

REFERENCES

1. Zimmerman JE, Kramer AA, McNair DS, et al: Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34:1297–1310
2. Le Gall JR, Neumann A, Hemery F, et al: Mortality prediction using SAPS II: An update for French intensive care units. *Crit Care* 2005; 9:R645–R652
3. Moreno RP, Metnitz PG, Almeida E, et al; SAPS 3 Investigators: SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005; 31:1345–1355
4. Knaus WA, Draper EA, Wagner DP, et al: APACHE II: A severity of disease classification system. *Crit Care Med* 1985; 13:818–829
5. Derde LPG, Cooper BS, Goossens H, et al; MOSAR WP3 Study Team: Interventions to reduce colonisation and transmission of antimicrobial-resistant bacteria in intensive care units: An interrupted time series study and cluster randomised trial. *Lancet Infect Dis* 2014; 14:31–39
6. van Duijn PJ, Verbrugghe W, Jorens PG, et al: The effects of antibiotic cycling and mixing on antibiotic resistance in intensive care units: A cluster-randomised crossover trial. *Lancet Infect Dis* 2018; 18:401–409
7. Plantinga NL, de Smet AMG, Oostdijk EA et al: Selective digestive and oropharyngeal decontamination in medical and surgical ICU patients: An individual patient data meta-analysis. *Clin Microbiol Infect* 2018; 24:505–513
8. Keegan MT, Gajic O, Afessa B: Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance. *Chest* 2012; 142:851–858
9. Brinkman S, Bakhshi-Raiez F, Abu-Hanna A, et al: External validation of Acute Physiology and Chronic Health Evaluation IV in Dutch intensive care units and comparison with Acute Physiology and Chronic Health Evaluation II and Simplified Acute Physiology Score II. *J Crit Care* 2011; 26:105.e11–e18
10. Lee H, Shon YJ, Kim H, et al: Validation of the APACHE IV model and its comparison with the APACHE II, SAPS 3, and Korean SAPS 3 models for the prediction of hospital mortality in a Korean surgical intensive care unit. *Korean J Anesthesiol* 2014; 67:115–122
11. van de Klundert N, Holman R, Dongelmans DA, et al: Data resource profile: The Dutch National Intensive Care Evaluation (NICE) Registry of Admissions to Adult Intensive Care Units. *Int J Epidemiol* 2015; 44:1850–1850h
12. Oostdijk EAN, Kesecioglu J, Schultz MJ, et al: Notice of retraction and replacement: Oostdijk *et al*. Effects of decontamination of the oropharynx and intestinal tract on antibiotic resistance in ICUs: A randomized clinical trial. JAMA. 2014;312(14):1429-1437. *JAMA* 2017; 317:1583–1584
13. Simonis FD, de Iudicibus G, Cremer OL, et al; MARS Consortium: Macrolide therapy is associated with reduced mortality in acute respiratory distress syndrome (ARDS) patients. *Ann Transl Med* 2018; 6:24
14. Stichting NICE Jaarboek 2016. Available at: https://www.stichting-nice.nl/doc/jaarboek-2016-web.pdf. Accessed January 1, 2019
15. Zimmerman JE, Kramer AA: Outcome prediction in critical care: The Acute Physiology and Chronic Health Evaluation models. *Curr Opin Crit Care* 2008; 14:491–497
16. Kramer AA, Zimmerman JE: Predicting outcomes for cardiac surgery patients after intensive care unit admission. *Semin Cardiothorac Vasc Anesth* 2008; 12:175–183
17. Tripepi G, Jager KJ, Dekker FW, et al: Stratification for confounding–part 1: The Mantel-Haenszel formula. *Nephron Clin Pract* 2010; 116:c317–c321
18. Debray TP, Moons KG, van Valkenhoef G, et al; GetReal Methods Review Group: Get real in individual participant data (IPD) meta-analysis: A review of the methodology. *Res Synth Methods* 2015; 6:293–309
19. Moons KG, Donders RA, Stijnen T, et al: Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006; 59:1092–1101
20. Sterne JA, White IR, Carlin JB, et al: Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* 2009; 338:b2393
21. Donders AR, van der Heijden GJ, Stijnen T, et al: Review: A gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006; 59:1087–1091
22. Van Buuren S, Groothuis-Oudshoon K: MICE: Multivariate Imputation by Chained Equations in R. *J Stat Softw* 2010; 45:1–68
23. Rubin DB: Multiple Imputation for Nonresponse in Surveys. New York, NY, John Wiley & Sons, 1987
24. Rosner B, Spiegelman D, Willett WC: Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *Am J Epidemiol* 1990; 132:734–745