ORIGINAL ARTICLE

Indirect and direct standardization for evaluating transplant centers

Kevin He*

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, United States

Received: October 14, 2018 **Accepted:** November 26, 2018 **Online Published:** December 10, 2018
DOI: 10.5430/jha.v8n1p9 **URL:** https://doi.org/10.5430/jha.v8n1p9

ABSTRACT

To assess the quality of health care, patient outcomes associated with medical providers are routinely monitored in order to identify poor (or excellent) provider performance. To avoid confounding by risk factors, both indirect and direct standardization have been used for comparing outcome rates or prevalence for different providers. There has been an ongoing debate as to which standardization method is more appropriate. To compare the performance of indirect and direct standardization for the purpose of ranking transplant centers, we analyzed post-transplant mortality using the national kidney transplant data. Included in our analysis were 116,601 patients (from 230 transplant centers) who underwent kidney transplantation between January 2006 and December 2012. Multivariate logistic regression model was used to model the 30-day mortality, which were estimates of failures (grant failure or death) in the 30 days after the transplant surgery. Concordance indexes, kappa coefficients and Spearman's rank correlation coefficient were computed. The estimated values from these statistics for the indirect standardized method were similar to the direct standardization. The results suggest that both indirect and direct standardized methods provide similar ability to distinguish center effects.

Key Words: Concordance index, Direct standardization, Indirect standardization, Kappa coefficient, Spearman's rank correlation, Transplant center

1. INTRODUCTION

Monitoring outcomes of medical providers is an important activity that has received much attention.[1–4] In order to identify poor (or excellent) performance of medical providers, outcomes of patients associated with providers are routinely monitored. This monitoring can help patients make more informed decisions, and can also aid consumers, stakeholders, and payers in classifying providers based on their performance. Given the high stakes of such evaluations, it is important that the methods used for profiling providers are appropriate.

Our endeavor here is motivated by the study of end-stage renal disease (ESRD), one of the most deadly and costly diseases in the United States. On December 31, 2015, there were 703,243 prevalent cases of ESRD in the U.S., which represents an increase of 80% since 2000.[5] Kidney transplantation is the preferred treatment for ESRD. However, despite aggressive efforts to increase the number of kidney donors, the demand far exceeds the supply. In 2015, the kidney transplant waiting list had 83,978 candidates, with fewer than 16% of eligible patients likely to receive a transplant.[5] At the same time, transplant patients often differ in response. To optimize the survival benefit of transplantations, one would anticipate significant benefit from a comprehensive care monitoring system.

*Correspondence: Kevin He; Email: kevinhe@umich.edu; Address: Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, United States.

The evaluation of center-specific mortality rates is a commonly used strategy in studies of post-transplant mortality among kidney transplant patients, with a view to ultimately improving patient care. It is well known that comparisons of unadjusted results (e.g., rates or ratios) can be misleading, since patient characteristics vary considerably across centers. In other words, factors other than center effects (such as socio-economic factors and comorbidities) could be responsible for any corresponding differences in center-specific crude results. Therefore, an accurate evaluation of center performance needs to account as much as possible for these confounding effects.

The desire to make valid (unconfounded) comparisons across centers motivates the use of risk-adjusted standardized methods. Generally, one of two standardized methods is used to examine the center effects on quality outcomes – indirect standardization and direct standardization. Indirect standardization has been widely used in epidemiology and sociology to compare event rates among.[6–9] In the field of provider profiling, various quality measure based on indirect standardizations have been proposed.[3, 10–13] The indirect standardized measure is designed to summarize the events at a center relative to the events that would be expected, based on the characteristics of the patients at that center, typically computed using rates estimated through a regression model. Results obtained through indirect standardization are often reported as ratios. For example, the method equals the ratio of the actual number of events divided by the expected number of events, where the latter is computed under a population norm. In the setting of interest (profiling transplant centers), the objective is essentially to identify outlier treatment centers. For this purpose, the population norm is estimated by pooling all the centers being compared in the observed sample. Qualitatively, the degree to which the center's indirect standardized measure varies from 1.00 is the degree to which it exceeds (> 1.00) or is under (< 1.00) the national event rates for patients with the same characteristics as those in the center. Although indirect standardized measure is a commonly used measure for internal evaluation (e.g., for centers to evaluate themselves or for a governing body to evaluate this center's event rate comparing to that expected at the national level), it has been argued that sets of indirect standardized measures should not be compared with one another, since each center-specific estimator is essentially adjusted using different covariate distributions.[4, 14]

Alternatively, direct standardization has also been applied in evaluations of medical providers.[14] One can express the direct standardized measure as the ratio of expected to observed numbers of events in the whole study population; the numerator of the direct standardized measure represents the expected number of events if all patients were treated at the given center, while the denominator equals the total observed number of events in the study population (across all centers). This way, the centers being compared are all averaged across the same adjustment covariate distribution. Thus, the same standard population is applied to all centers, and hence, direct standardization are directly comparable. Despite its advantage, the direct standardization may not be easily understood by the investigators or other stakeholders, with investigators often chiefly interested in quantities more directly comparing center-specific outcomes with the population norm. In contrast, the indirect standardization has a long history in fields of profiling medical providers, and it is a valid metric for internal evaluation. Moreover, another disadvantage of direct standardization is the implicit requirement that adjustment covariate-specific rates be sufficiently precise for each center being compared. For settings where the event of interest is rare, this is often not the case. This issue may explain the wide uses of indirect standardization, which requires sufficiently precise rates at the population (as opposed to individual center) level.

There has been an ongoing debate as to which standardization method is more appropriate. To investigate this problem, we used national kidney transplant data obtained from U.S. Organ Procurement and Transplantation Network (OPTN) to examine 30-day post-transplant mortality. We performed simulations based on the real data cohort to compare the performance of indirect and direct standardization in terms of ranking transplant centers. Our paper continues as follows: Section 2 introduces the data source and the quality measure used to assess the indirect and direct standardization. Section 3 examines the indirect and direct standardization with simulations using national data on kidney transplant patients. We conclude with a discussion in Section 4.

2. Methods

2.1 Data sources and variables

Data for this study were Data were obtained from the U.S. OPTN. Included in our analysis were 116,601 patients (from 230 transplant centers) who underwent kidney transplantation between January 2006 and December 2012. Patient failure time (recorded in days) was defined as the time from transplantation to graft failure or death, whichever occurred first. Adjustment covariates (p = 25) in this study included baseline recipient characteristics such as age at transplantation, race, gender, BMI, indicator of previous kidney transplant, and comorbidity conditions (e.g. polycystic kidney disease, diabetes, IgA nephropathy and malignant tumor), and donor characteristics such as cold ischemic time, type of donor kidney. Race was categorized as White, African

American, Hispanic, Asian or other. Cold ischemia times were categorized as Low (20 hours or less) or High (longer than 20 hours). Type of donor kidney was categorized as living, standard criteria donor, or expanded criteria donor (ECD).

The main outcome considered in this study was the 30-day mortality, which were estimates of failures (grant failure or death) in the 30 days after the transplant surgery. The mortality rates were measured within 30 days, because deaths after a longer time period may have less to do with the care provided in the transplant centers and more to do with other complicating illness, patient's own behaviors, or care provided to patients out of control of transplant centers.

2.2 Statistical model

We consider a multivariate logistic regression model in which centers are represented as fixed effects. Let Y_{ij} denote the observed outcome for the jth observation within the ith center, where *i* = 1, 2, . . . , *F* and *j* = 1, 2, . . . , n_i, with n_i being the number of observations in center *i*. In the context of our cohort, Y$_{ij}$ equals 1 if the jth transplant in the ith center results in a death or grant failure within 30 days, and Y$_{ij}$ equals 0 otherwise. The corresponding logistic regression is of the form (see Equation 1):

$$logit(p_{ij}) = log\frac{p_{ij}}{1-p_{ij}} = \alpha_i + X_{ij}^T\beta \qquad (1)$$

where p_{ij} is the probability that Y_{ij} equals 1, α_i corresponds to the fixed center effects, X_{ij} is the covariates associated with the observation, and β is the regression parameter. In our cases, α_i measures the center effect in the sense that a large value of α_i would indicate that the ith center performs relatively poorly.

2.3 Indirect standardization

Under the logistic regression model, an indirect standardized ratio (ISR) for the ith center can be stipulated as Equation 2:

$$ISR_i = \frac{O_i}{E_i} \qquad (2)$$

where O_i = $\sum_{j=1}^{n_i} Y_{ij}$ is the observed number of events in center *i*, and

$$\sum_{j=1}^{n_i} p_{ij}(\hat{\alpha}_M, \hat{\beta}) = \sum_{j=1}^{n_i} \frac{exp(\hat{\alpha}_M + X_{ij}^T\hat{\beta})}{1+exp(\hat{\alpha}_M + X_{ij}^T\hat{\beta})} \qquad (3)$$

Equation 3 is the expected number. The latter is the sum of the estimated probabilities of all observations within this center, assuming a national norm for the center effect, which is specified with $\hat{\alpha}_M$ = median ($\hat{\alpha}_1$, . . . , $\hat{\alpha}_F$), e.g., the median of the estimated center effects. Note that we adopt a median term for the "average" center effect; this is more robust to extreme values and avoids problems that would arise in using the mean.[2] Here $\hat{\beta}$ is the estimated regression parameters for patient characteristics. In this ratio, each center is compared with an average center, adjusting for its particular patient characteristics. An ISR lower than 1 indicates that the center's observed event rate is less than expected based on national rates and vice versa.

2.4 Direct standardization

Alternatively, a direct standardized ratio (DSR) for the ith center is given by Equation 4:

$$DSR_i = \frac{\sum_{k=1}^{F}\sum_{j=1}^{n_k} exp(\hat{a}_i + X_{kj}^T\hat{\beta})}{\sum_{k=1}^{F} O_k} \qquad (4)$$

where the denominator is the total number of observed events across all centers, and the numerator is the total number of expected events across all centers, assuming all centers have event rates equal to that of center *i*. Thus, the DSR also involves a ratio of observed and expected numbers of events. However the expected component is in the DSR's numerator, while the observed count is in the denominator. With respect to interpretation, a DSR greater than 1 indicates that this center has a greater event rate than the overall average.

2.5 Center ranking

To assess the ability of ISR and DSR to distinguish between centers, we considered three statistical metrics. The first one was a concordance index for center ranking. For ISR, the concordance index was given by Equation 5:

$$C_{ISR} = \frac{\sum_{1\leq i\leq k\leq F}(ISR_i \leq ISR_k, \alpha_i \leq \alpha_k)}{\sum_{1\leq i\leq k\leq F}(ISR_i \leq ISR_k)} \qquad (5)$$

which was the proportion to which the ranking based on ISR was consistent with the true ranking. Here α_i and α_k were the true effects for center *i* and *k*. A large value of C_{ISR} indicates that the method has a better ability to distinguish center effects. Similarly, for DSR, the concordance index was given by Equation 6:

$$C_{DSR} = \frac{\sum_{1\leq i\leq k\leq F}(DSR_i \leq DSR_k, \alpha_i \leq \alpha_k)}{\sum_{1\leq i\leq k\leq F}(DSR_i \leq DSR_k)} \qquad (6)$$

To further assess the agreement for distinguishing centers, we computed Cohen's kappa (κ) coefficients[15] for both ISR and DSR. Specifically, centers were categorized into three equal-sized groups based on either their true center effects or the estimated standardized ratios. Then the Cohen's kappa was computed to measure agreement between two raters. In particular, the Cohen's kappa coefficients for ISR and DSR were computed using categories based on true effects versus those categories based on ISR and DSR, respectively. A larger value of kappa coefficient indicates a better agreement between the rater based on the standardized method and the one based on true center effects. Finally, Spearman's rank correlation coefficient, which measures statistical dependence between rankings, was also computed to assess how well the relationship between the true ranking and those based on either indirect or direct standardized methods can be described using a monotonic function. Intuitively, the Spearman correlation between two rankings will be high when observations have a similar rank, and low when observations have a rank between the two variables.

3. Results

Over the period from 2006 to 2012, among 230 centers placing at least 50 kidney transplants, there were a total of 116,601 transplants, yielding an overall 30-day mortality rate of 2.3%. The number of transplant centers per center varied from 50 to 2,238, with a mean of 507.0 and a median of 362.5 transplants. Table 1 presents a multivariate logistic regression model for the outcome of 30-day mortality. Of patient characteristics, the odds of 30-day mortality were lower with Asian recipient race, male recipient sex, living donor type and the presence of recipient comorbidities such as Polycystic Kidney Disease. In contrast, the odds of 30-day mortality were higher for recipient with age less than 10, over-weight or obesity, the presence of diabetes, previous kidney transplant and high cold ischemia times.

Table 1. Patient characteristics as predictors of outcomes of 30-day mortality

	OR	**95% CI**		***p*-value**
Recipient Age, years (ref: 25-34)				
• ≤ 10	2.34	1.64	3.34	< .001
• 11-17	1.26	0.93	1.70	.136
• 18-24	1.08	0.83	1.42	.554
• 35-44	1.03	0.87	1.23	.699
• 45-54	1.04	0.89	1.23	.612
• 55-64	1.12	0.96	1.32	.154
• 65-74	1.18	0.99	1.41	.057
• ≥ 75	1.15	0.84	1.55	.383
Recipient Race (ref: White)				
• Asian	0.76	0.62	0.94	.010
• Black	1.05	0.95	1.16	.318
• Other/Unk/Miss	1.04	0.75	1.43	.832
Recipient Ethnicity				
• Hispanic	0.93	0.81	1.06	.256
Recipient Gender (ref: Female)				
• Male	0.86	0.79	0.93	.002
Recipient BMI (ref: normal)				
• Under (< 18.5)	1.11	0.88	1.40	.371
• Over (25.5-30)	1.27	1.15	1.40	< .001
• Obesity (> 30)	1.32	1.19	1.47	< .001
Recipient Comorbidities (ref: no)				
• Diabetes	1.41	1.16	1.72	.001
• Polycystic Kidney Disease	0.76	0.65	0.89	.001
• IgA nephropathy	0.81	0.65	1.01	.056
• Malignant Tumor	1.06	0.89	1.26	.500
Prior Kidney Transplant	1.28	1.15	1.44	< .001
Donor Type (ref: Deceased)				
• Living	0.52	0.47	0.58	< .001
Donor Gender (ref: Female)				
• Male	0.98	0.91	1.06	.643
Expanded Criteria Donor	1.61	1.45	1.79	< .001
Cold Ischemia Times (ref: Low)				
• High (longer than 20 hours)	1.14	1.03	1.25	.009

Note. Multivariate logistic regressions were implements for 30-day mortality. Included in our analysis were 49,142 patients (from 200 transplant centers) who underwent kidney transplantation between January 2010 and December 2012

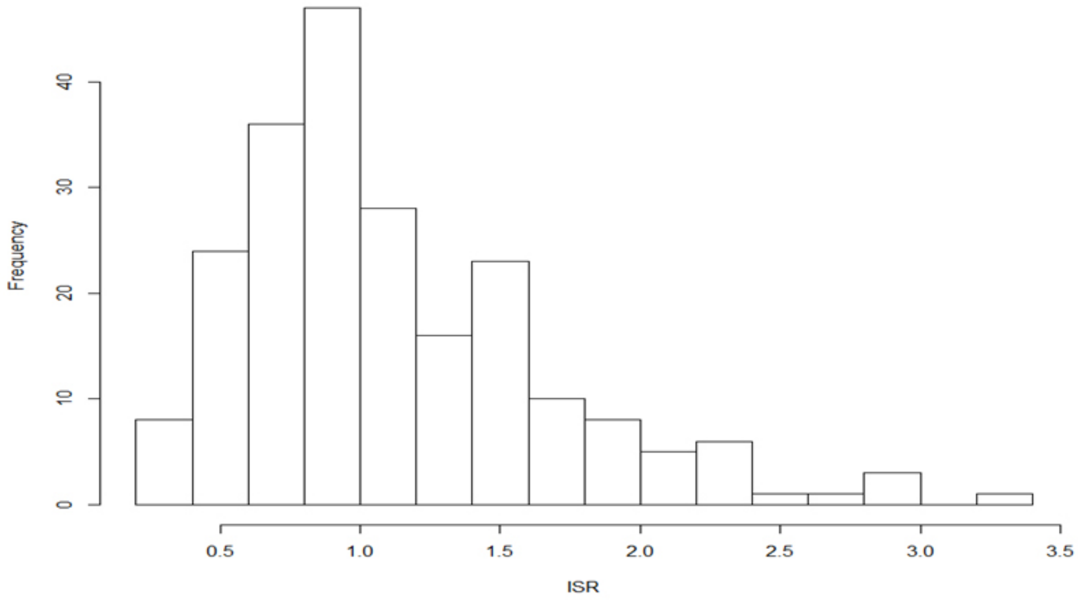

Figure 1. Histogram of estimated ISR, limited to centers with at least 50 transplant centers over the study period

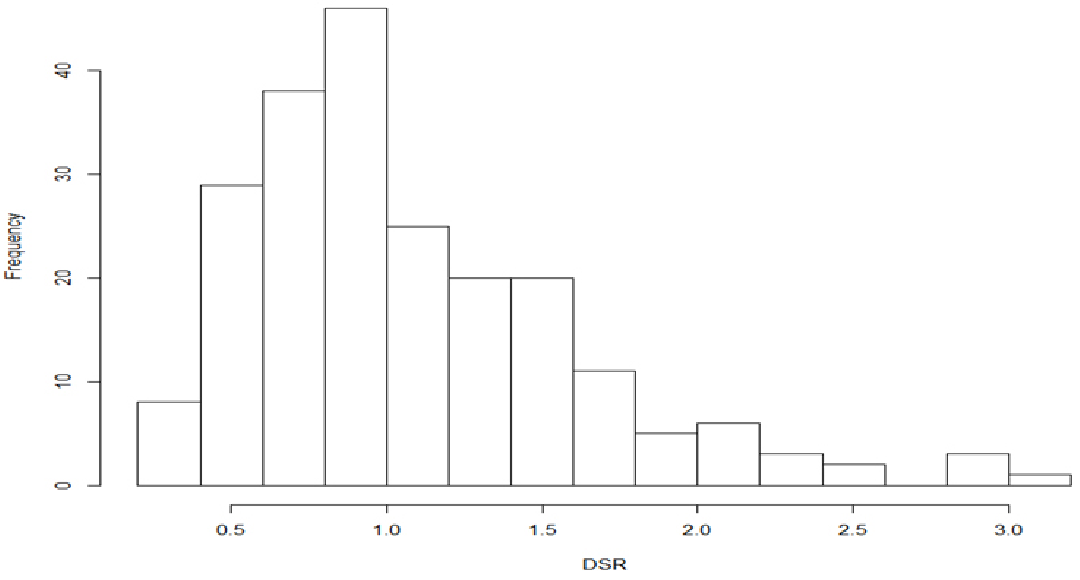

Figure 2. Histogram of estimated DSR, limited to centers with at least 50 transplant centers over the study period

Figures 1 and 2 show the histograms of the estimated center-specific ISR and DSR. There were wide variations among transplant centers. Figure 3 shows the scatterplot and presents the pairwise comparisons of the ISR and DSR. Figure 4 shows the boxplot and compares the distributions of these two standardized methods. In this settings, both methods provide similar center effect estimates.

In order to further examine the abilities of the standardized methods to distinguish center effects, we carried out simulations based on the real data characteristics. The observed outcomes were generated from a multivariate logistic regression as in (1), where the data structure (e.g., number of centers and number of observation within each center) and patient characteristics were the same as those in the real data cohort. Moreover, the true center effects and covariate effects were chosen as the estimated values from the real data. As shown in Table 2, the estimated concordance indexes, kappa coefficients and Spearman's rank correlation coefficients for the indirect standardized method were similar to the direct standardization across all setting.

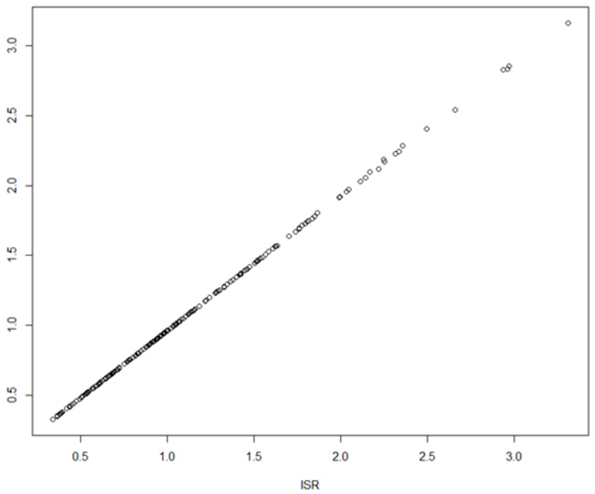

Figure 3. Scatterplot of estimated ISR and DSR, limited to centers with at least 50 transplant centers over the study period

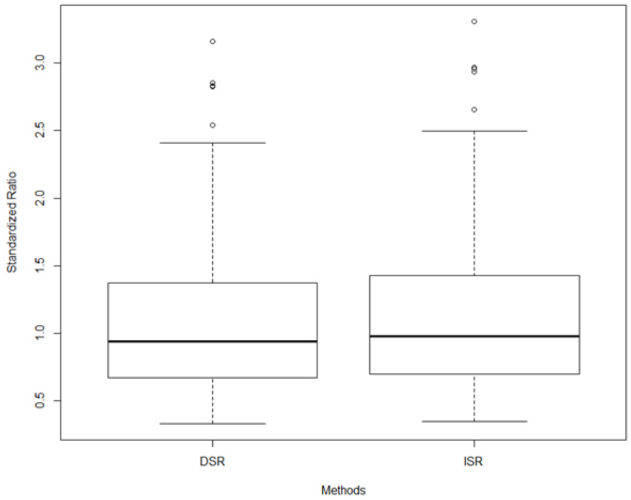

Figure 4. Boxplot of estimated ISR and DSR, limited to centers with at least 50 transplant centers over the study period

Table 2. Patient characteristics as predictors of outcomes of 30-day mortality

Method	Concordance Index	Cohen's Kappa Coefficient	Spearman's Rank Correlation
Indirect	0.8006	0.7231	0.7830
Direct	0.8006	0.7231	0.7830

Note. The Concordance Index was computed as the proportion that the pairwise ranking based on ISR or DSR was consistent with the true ranking. The Cohen's Kappa Coefficient was computed to measure agreement between ordered tertiles based on true effects versus tertiles based on ISR or DSR

4. Discussion

To assess the quality of health care, patient outcomes associated with transplant centers are routinely monitored in order to identify poor (or excellent) center performance. To avoid confounding by risk factors, both indirect and direct standardization have been used for comparing outcome rates or prevalence for different centers. In particular, we summarized the advantages and drawbacks of using indirect standardization for profiling providers:

(1) Indirectly standardized estimators provides a valid approach to evaluate how does a center's event rate or prevalence compare to that predicted at the population level.

(2) The use of indirect standardization only requires sufficiently precise rates at the population (as opposed to an individual center) level.

(3) The implementation of indirect standardization is straightforward and the interpretation the indirect standardization is relatively easily understood by the investigators or other stakeholders.

(4) However, it has been argued that the center-specific indirect standardized estimators may not be compared with one another, because each center's indirect standardized measure is essentially adjusted to a different (center-specific) covariate distribution.

An alternative approach is to adjust all measures to the overall distribution of covariates combining across all centers, for which we summarize the corresponding advantages and drawbacks as follows:

(1) The direct standardization makes all measures comparable to each other.

(2) However, a disadvantage of direct standardization is the implicit requirement that the rates be sufficiently precise for each center, which can be violated in practice.

(3) The direct standardization may not be easily understood by the investigators or other stakeholders.

In practice, there is often some judgement required in deciding when to use indirect standardization and when to use direct standardization. After analyzing the 30-day mortality for kidney transplant patients, using the national kidney transplant database from 2006-2012, the results based on concordance indexes, kappa coefficients and Spearman's rank correlation coefficients suggest that at least in our settings the indirect standardized method provides similar ability as the direct standardized approach to distinguish center effects.

In conclusion, the rationale for the argument that direct standardization is better than the indirect standardization to rank-order medical providers is based on the assumption that adjustment covariate-specific rates can be sufficiently precise for each provider being compared. For settings where the event of interest is rare or the center-specific sample sizes are small, this assumption is often violated. Our results confirm the advantage of the wide uses of indirect standardization, which requires sufficiently precise rates at the population (as opposed to individual center) level. Thus, in monitoring of transplant centers, we prefer indirect standardization as it is easy to compute and interpret, giving a more meaningful measure to each center comparing their results with the national norm for the patients they actually treat.

ACKNOWLEDGEMENTS

The authors thank Dr. Kirsten Herold at the UM-SPH Writing Lab for her helpful suggestions.

CONFLICTS OF INTEREST DISCLOSURE

The authors declare they have no conflicts of interest.

REFERENCES

[1] Liu D, Schaubel DE, Kalbeisch JD. Computationally efficient marginal models for clustered recurrent event data. Biometrics. 2012; 68: 637-647. PMid: 21957989. https://doi.org/10.1111/j.1541-0420.2011.01676.x

[2] He K, Kalbeisch JD, Li Y, et al. Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. Lifetime Data Analysis. 2013; 19(4): 490-512. PMid: 23709309. https://doi.org/10.1007/s10985-013-9264-6

[3] Kalbeisch JD, Wolfe RA. On monitoring outcomes of medical providers. Statistics in Biosciences. 2013; 5(2): 286-302. https://doi.org/10.1007/s12561-013-9093-x

[4] Estes JP, Nguyen DV, Chen Y, et al. Time-dynamic profiling with application to hospital readmission among patients on dialysis. Biometrics. 2018.

[5] Saran R, Robinson B, Abbott KC, et al. US Renal Data System 2016 Annual Data Report: Epidemiology of Kidney Disease in the United States. American journal of kidney diseases: the official journal of the National Kidney Foundation. 2017; 69(3 Suppl 1): A7-A8. PMid: 28236831. https://doi.org/10.1053/j.ajkd.2016.12.004

[6] Berry G. The analysis of mortality by the subject-years method. Biometrics.1983; 39: 173-184. PMid: 6871346. https://doi.org/10.2307/2530817

[7] Breslow NE, Day NE. The standardized mortality ratio. Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences. The Bernard G. Greenberg Volume Edited by P.K. Sen. 1985; 55-74.

[8] Keiding N. The method of expected number of deaths, 1786-1886-1986. International Statistical Review. 1987; 55: 1-20. PMid: 12179583. https://doi.org/10.2307/1403267

[9] Hazel I. Standardization methods. Encyclopedia of Biostatistics. 2nd ed. 2005; 5151-5163.

[10] Wolfe RA, Gaylin DS, Port FK, et al. Using USRDS generated mortality tables to compare local ESRD mortality rates to national rates. Kidney International. 1992; 42: 991-996. PMid: 1453592. https://doi.org/10.1038/ki.1992.378

[11] Wolfe RA. The standardized morality ratio revisited: improvements, innovations, and limitations. American Journal of Kidney Diseases. 1994; 24: 290-297. https://doi.org/10.1016/S0272-6386(12)80194-6

[12] Dickinson DM, Shearon TH, O'Keefe J, et al. SRTR center-specific reporting tools: post transplant outcomes. American Journal of Transplantation. 2006; 6: 1198-1211. PMid: 16613596. https://doi.org/10.1111/j.1600-6143.2006.01275.x

[13] Ash AS, Fienberg SE, Louis TA, et al. Statistical issues in assessing hospital performance. The COPSS-CMS White Paper. 2011.

[14] He K, Schaubel DE. Standardized mortality ratio for evaluating center-specific mortality: assessment and alternative. Statistics in Bioscience. 2014; 7(2): 296-321. https://doi.org/10.1007/s12561-014-9119-z

[15] Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20(1): 37-46. https://doi.org/10.1177/001316446002000104