ORIGINAL ARTICLE

Inter-unit reliability for quality measure testing

Kevin He*, John D. Kalbfleisch, Yuan Yang, Zhe Fei, Sehee Kim, Jiang Kang, Yi Li

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

Received: October 21, 2018 **Accepted:** January 4, 2019 **Online Published:** January 8, 2019
DOI: 10.5430/jha.v8n2p1 **URL:** https://doi.org/10.5430/jha.v8n2p1

ABSTRACT

Facility-specific quality measures are commonly used to monitor dialysis facilities. To successfully develop, test and validate quality measures, a subset of facilities are often recruited for preliminary evaluations. To ensure that the facility-specific measures will achieve a desirable precision, it is often of interest to determine a minimum number of facilities that should be recruited. To achieve this, we propose a method based on the inter-unit reliability (IUR), which is commonly used to assess quality measures. Accurate estimates of the IUR are important to ensure that the quality measure will achieve a desirable precision. We first review existing methods of estimating the IUR for quality measures that are based on sample averages. We then generalize the IUR estimations to more complicated standardized measures. In particular, the confidence intervals of the IUR are calculated, with the width of this confidence interval measuring the precision of the estimate of the IUR. To assess the performance of the estimated IUR with various numbers of facilities, a simulation study is conducted. The IURs are then computed to develop and implement a quality measure that is used to guard against high ultrafiltration rates for adult dialysis patient with End-Stage Renal Disease. The estimated values are helpful to determine a minimum number of facilities that should be recruited in the measure testing process.

Key Words: IUR, Medical profiling, Power, Reliability, UFR

1. INTRODUCTION

Multi-provider (e.g. hospital, transplant center or dialysis facility) data arise very often in biomedical studies. In order to identify extreme (excellent or poor) performance, provider-specific clinical outcomes are routinely monitored.[1–3]

The motivating example is to develop and implement quality measures that will be used to guard against high ultrafiltration (i.e., rapid fluid removal) rates for adult dialysis patient with End-Stage Renal Disease (ESRD). ESRD is one of the most deadly and costly diseases in the United States. As of December 31, 2015, there were 703,243 prevalent cases of ESRD in the United States, which represents an increase of 80% since 2000.[4] Hemodialysis (HD) is the most commonly used treatment option for patients with ESRD. However, despite the majority of dialysis patients achieving urea removal, the mortality rate among HD patients is substantially higher than the general population.[4] Existing literature suggests that higher ultrafiltration rate (UFR) is an important predictor of mortality.[5–7] The UFR measures the rapidity with which fluid is removed at dialysis per unit (kg) body weight in unit (hour) time. The UFR is under control of the dialysis facilities. Rapid UFR can lead to higher frequency of intradialytic hypotension, which occurs at high frequency and has been associated with higher mortality. Phenomena, such as myocardial stunning,[8,9] could result if large volumes of fluid are removed rapidly during each dialysis session.

This report is intended to develop a safety measure to guard against high UFR. To successfully develop such quality measures for dialysis facilities, measure testing is typically conducted.[2,3] However, due to access feasibility and funding

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

limitations in the measure testing process, it is not always possible to enroll all (more than 5,000) dialysis facilities across the United States. As a result, a subset of facilities are often selected for preliminary evaluations. To determine a minimum number of facilities to ensure a desirable precision, accurate estimates of the inter-unit reliability (IUR) are important.

Statistically, the IUR of a facility-specific measure is defined as the ratio of the between-facility variance to the total variation,[10, 11] where the total variation equals the sum of the between-facility variance and the within-facility variance. IUR values range between 0 and 1. A small IUR indicates that most of the variation in the facility-specific quality measure can be regarded as random noise. A large IUR indicates that most of the variation is due to real differences between facilities. Accurate estimates of the IUR are important to ensure that the quality measure will achieve a desirable precision.

The report continues as follows: Section 2 introduces the motivating quality measure and describes the methods used to estimate the IUR. Finite-sample properties are examined in Section 3 through simulation studies. Section 4 summarizes the analysis results on the motivating example data. We conclude with a discussion in Section 5.

2. METHOD

2.1 Motivating example

Our motivating example for quality measures is to evaluate percentage of adult kidney hemodialysis patients with a monthly ultrafiltration rate (UFR) greater than 13. Higher dialytic ultrafiltration rates have been shown to be associated with higher all-cause and cardiovascular mortality.[7] The UFR is calculated based on the weight gain after the dialysis session and the delivered session time. It is well known that hemodialysis patients have high rates of morbidity and mortality that may be related to rapid ultrafiltration. Thus, a quality measure based on the ultrafiltration rate could encourage dialysis facilities to consider prolonging treatment time and ensure stable dialysis sessions.

2.2 IUR for UFR without risk adjustment

Methods for estimating the IUR vary by whether adjustment for risk factors are needed. We first provide a brief review of IUR for UFR without adjustment for risk factors. In next subsection, we then generalize the IUR for more complicated risk-adjusted quality measures.

We first consider a logistic regression model without adjustment of risk factor. Let Y_{ij} denote the observed outcome (e.g. indicator for whether UFR greater 13) for patient j in facility i, where $j = 1, \ldots, n_i$ and $i = 1, \ldots, F$, with n_i being the number of patients in facility i and F being the number of facilities. Here $Y_{ij} = 1$ if the UFR rate is greater 13, and $Y_{ij} = 0$ otherwise. The corresponding logistic regression is of the form (see Equation 1):

$$\text{logit}(p_{ij}) = \log \frac{p_{ij}}{1 - p_{ij}} = \mu + \alpha_i \tag{1}$$

where μ is the intercept for the population norm, $p_{ij} = P(Y_{ij} = 1|\alpha_i)$, and $\alpha_i \sim N(0, \sigma_\alpha^2)$ is the facility effect. In this case, the facility-specific measure is defined as the sample average of observed outcomes across facility, e.g., $\overline{Y}_i = \sum_{j=1}^{n_i} Y_{ij}$.

To compute the IUR for the facility-specific measure, a key ingredient is to decompose the variance of the measure into the between and within-facility variations. To achieve this, we adopt the following variance decomposition (see Equation 2):

$$\text{Var}(\overline{Y}_i) = \text{Var}(\text{E}(\overline{Y}_i|\alpha_i)) + \text{E}(\text{Var}(\overline{Y}_i|\alpha_i)) \tag{2}$$

Applying a first-order Taylor expansion, the following approximations can be computed (see Equation 3):

$$\begin{aligned} &\widehat{\text{Var}}(\text{E}(\overline{Y}_i|\alpha_i)) = \sigma_\alpha^2\{\overline{p}(1-\overline{p})\}^2 \\ &\text{and } \text{E}(\text{Var}(\overline{Y}_i|\alpha_i)) = \overline{p}(1-\overline{p})/n_i \end{aligned} \tag{3}$$

where $\overline{p}$ is the sample average of Y_{ij} in the overall study sample. The corresponding IUR is given by (see Equation 4):

$$\begin{aligned} IUR &= \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + 1/\{n'\overline{p}(1-\overline{p})\}} \\ \text{where } n' &= \frac{1}{F-1}\left(\sum_{i=1}^{F} n_i - \frac{\sum_{i=1}^{F} n_i^2}{\sum_{i=1}^{F} n_i}\right) \end{aligned} \tag{4}$$

We note that a similar deviation was considered by Yelland et al.[12] to estimate the intra-class correlation coefficients and calculate appropriate sample size for clustered clinical trials. There the number of clusters was fixed and the main goal of the sample size calculation was to determine an appropriate number of patients to be enrolled. In contrast, in our motivating setting, the main goal is to determine an appropriate number of facilities.

2.3 IUR for UFR with risk adjustment

In practice, patient characteristics may vary across facilities and a large number of high-risk patients can make a facility's outcomes appear substandard. To account for the covari-

ance imbalance, a risk-adjusted logistic regression can be implemented (see Equation 5):

$$\text{logit}(p_{ij}) = \log \frac{p_{ij}}{1-p_{ij}} = \mu + \alpha_i + X_{ij}^T\beta \quad (5)$$

where μ is an intercept for the population average, X_{ij} is a vector of covariates, β is a vector of regression parameters, $\alpha_i \sim N(0, \sigma_\alpha^2)$ is the facility effect, and $p_{ij} = P(Y_{ij} = 1|\alpha_i, X_{ij})$.

Under this logistic model, the standardized ratio (SR) is defined as (see Equation 6):

$$SR_i = \sum_{j=1}^{n_i} Y_{ij} / \sum_{j=1}^{n_i} \hat{p}_{ij}, \quad (6)$$

where the numerator $\sum_{j=1}^{n_i} Y_{ij}$ is the observed number of events in facility i, and the denominator $\sum_{j=1}^{n_i} \hat{p}_{ij}$ is the corresponding expected number, with (see Equation 7):

$$\hat{p}_{ij} = p_{ij}(\hat{\mu}, \widehat{\boldsymbol{\beta}}) = \frac{\exp(\hat{\mu} + X_{ij}^T\widehat{\boldsymbol{\beta}})}{1+\exp(\hat{\mu} + X_{ij}^T\widehat{\beta})} \quad (7)$$

Thus, the denominator is the sum of the estimated probabilities within each facility under a population norm for the facility effect, which is estimated from $\hat{\mu}$.

This standardized ratio reflects the extent to which the facility under evaluation has a higher or lower rate of events (e.g. UFR greater than 13) than the overall population average. A standardized ratio lower than 1 indicates that the facility's observed number of events is less than expected based on the national norm. A ratio greater than 1 indicates that the facility has a number of observed events higher than expected.

To compute the IUR of the standardized ratio, we exploit the large-scale structure of the data, which allows $\hat{\mu}$ and $\hat{\beta}$ to be estimated precisely when the sample size is large. Thus, we adopt similar techniques as in He et al.[2] and treat $\hat{\mu}$ and $\hat{\beta}$ as given without losing precision. Using the variance decomposition (see Equation 8):

$$\text{Var}(SR_i) = \text{Var}(\text{E}(SR_i|\alpha_i)) + \text{E}(\text{Var}(SR_i|\alpha_i)) \quad (8)$$

and applying a first-order Taylor expansion, the approximate marginal variance of SR_i is given by (see Equation 9):

$$\widehat{\text{Var}}(SR_i) = \frac{\sigma_\alpha^2\{\sum_{j=1}^{n_i}\hat{p}_{ij}(1-\hat{p}_{ij})\}^2}{(\sum_{j=1}^{n_i}\hat{p}_{ij})^2} + \frac{\sum_{j=1}^{n_i}\hat{p}_{ij}(1-\hat{p}_{ij})}{(\sum_{j=1}^{n_i}\hat{p}_{ij})^2} \quad (9)$$

The facility-specific IUR can be calculated as (see Equation 10):

$$IUR_i = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + 1/n_i^*} \quad (10)$$

where $n_i^* = \sum_{j=1}^{n_i}\hat{p}_{ij}(1-\hat{p}_{ij})$ is the effective sample size. The overall IUR across facilities can be computed as (see Equation 11):

$$IUR = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + 1/n^*}$$
$$\text{where } n^* = \frac{1}{F-1}\left(\sum_{i=1}^{F} n_i^* - \frac{\sum_{i=1}^{F}(n_i^*)^2}{\sum_{i=1}^{F} n_i^*}\right) \quad (11)$$

3. SIMULATION

3.1 IUR for UFR without risk adjustment

We first consider the setting without risk-adjustment. We simulate the outcomes from the logistic model defined as in Equation 1 with $\mu = 0.5$ and $\alpha_i \sim N(0, 0.3^2)$. We vary the numbers of facilities from 20 to 600. The numbers of patients within each facility are generated such that one third of facilities has 25, 50 and 100 patients, respectively. Figure 1 shows the average estimated IUR based on 100 simulation iterations and the confidence intervals. Clearly, the results indicate that small numbers of facilities are not sufficient for reliable estimation of IUR.

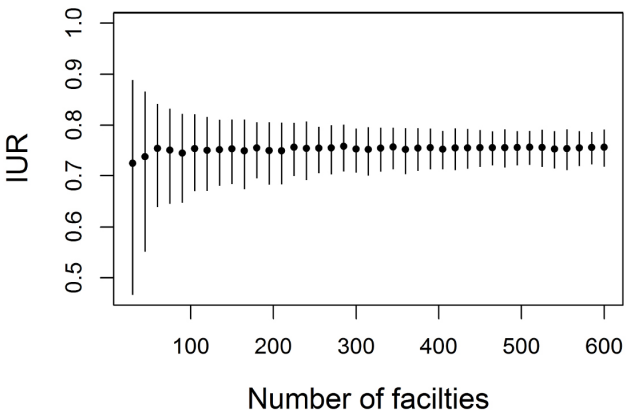

Figure 1. IUR for UFR without risk-adjustment

3.2 IUR for UFR with risk adjustment

To mimic scenarios with risk adjustment, we simulate a setting with the outcome Y_{ij} generated from a multivariate logistic model with probability (see Equation 12):

$$\text{logit}(p_{ij}) = \log \frac{p_{ij}}{1-p_{ij}} = \mu + \alpha_i + X_{ij1} - X_{ij2} \quad (12)$$

where $\alpha_i \sim N(0, 0.1^2)$, and X_{ij1} and X_{ij2} are generated

from independent standard normal distributions. We let the number of facilities vary from 20 to 600. The number of observations in each facility ranges from $120, 140, \ldots, 300$. We repeat the simulation 100 times. The average IUR and its empirical confidence interval are shown in Figure 2. Similar to the results in Section 3.1, a relatively large number of facilities is needed for reliable estimation of the IUR.

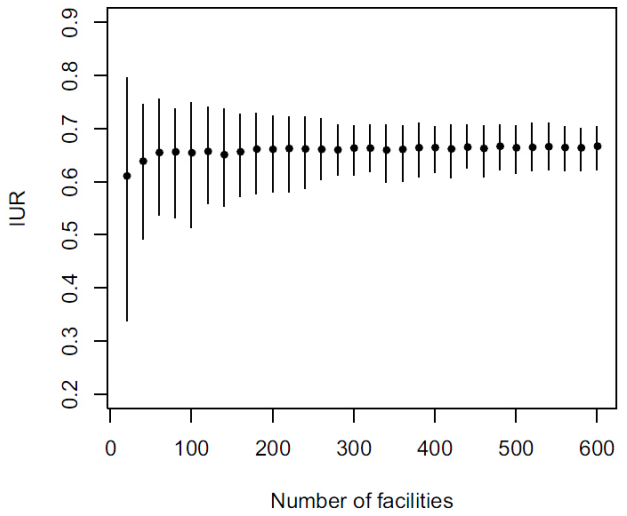

Figure 2. IUR for UFR with risk-adjustment

4. Data application

The goal of this section is to determine the minimum number of recruited facilities to ensure precise estimation of the overall IUR of a tested measure. We exemplify our procedure by using a dialysis adequacy measure, which is the percentage of adult hemodialysis patients (older than 18) in a facility with a monthly UFR greater than 13 ml/kg/hr. Patients with time since imitation of dialysis of less than 90 days, time within a facility of less than 90 days, missing weight information before or after dialysis session or missing delivered time of dialysis sessions are excluded from further analysis. Based on the 2012 CROWNWeb data, the average UFR is 9.2 with a median of 8.7 ml/kg/hr. The corresponding average percentage of UFR greater than 13 at the facility level is 19.3%. The total number of facilities is 5,317, with the 10th, 33th percentile, median, 67th, and 90th percentile for number of patients within each facility being 15, 35, 50, 68, and 108, respectively. The IUR based on the whole study population is 0.738. Figure 3 shows a histogram of the facility-level average UFR across a total of 5,317 facilities nationwide.

We randomly draw a certain number of facilities from the study population and calculate the IUR based on the method described in Section 2.2. This way, the variations of IURs are accounted for across different subsamples. To further improve the representativeness of the selected subsample, we consider facilities with various sizes (small, moderate or large).

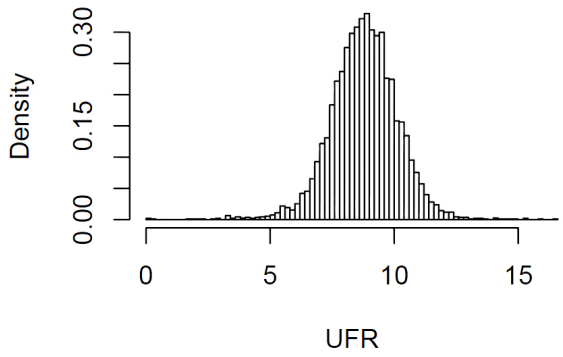

Figure 3. Histograms of UFR

Table 1. Average IUR and 95% CI for various numbers of facilities

Number of facilities	IUR (CI)
15	0.660 (0.015, 0.996)
18	0.714 (0.447, 0.885)
21	0.719 (0.454, 0.888)
24	0.719 (0.476, 0.878)
27	0.730 (0.528, 0.867)
30	0.730 (0.528, 0.868)
33	0.732 (0.545, 0.862)
36	0.733 (0.557, 0.857)
39	0.736 (0.564, 0.857)
42	0.739 (0.577, 0.855)
45	0.741 (0.585, 0.853)
48	0.741 (0.598, 0.845)
51	0.743 (0.612, 0.840)
54	0.744 (0.613, 0.843)
57	0.744 (0.614, 0.840)
60	0.745 (0.620, 0.840)
63	0.739 (0.607, 0.838)
66	0.743 (0.624, 0.834)
69	0.747 (0.633, 0.834)
72	0.746 (0.633, 0.833)
75	0.747 (0.636, 0.833)
78	0.748 (0.645, 0.830)
81	0.747 (0.644, 0.828)
84	0.742 (0.643, 0.821)
87	0.747 (0.649, 0.826)
90	0.749 (0.652, 0.826)

We then randomly draw facilities such that one third of the selected facilities are from the smallest tertile, one third are from the middle tertile, and one third are from the largest tertile. The confidence intervals of the IUR are calculated, using a logit transformation with the standard error of the logit IUR obtained from the Delta method. Here the logit transformation is applied to improve normality assumptions

and assure that the confidence interval is within the range of 0 and 1. We repeat this procedure 1,000 times and obtain the average confidence interval. The width of this average confidence interval measures the precision of the estimate of the IUR, and half the width estimates the corresponding average margin of error.

Table 1 shows the calculated average IURs and their confidence intervals against the number of facilities. It appears that wider confidence intervals are associated with smaller numbers of facilities, indicating unreliable estimation of IURs. Moreover, IURs tend to be under-estimated given a small number of facilities. To ensure the length of the confidence intervals is less than 0.2, at least 75 facilities are needed.

5. DISCUSSION

Facility-specific quality measures are commonly used to monitor dialysis facilities. To successfully develop, test and validate quality measures, a subset of facilities are often recruited for preliminary evaluations. To ensure that the facility-specific measures will achieve a desirable precision, the objective of this report is to determine a minimum number of facilities that should be recruited. To achieve this, we implement a method based on the IUR, which is commonly used to assess the reliability of quality measures for evaluating facilities. In general terms, the reliability of a measure reflects to what extent the measure reflects the actual differences between facilities as opposed to the random variation of patient outcomes within the facility. We calculate IURs and their confidence intervals for various numbers of facilities to determine a minimum number of facilities that should be recruited to ensure that the overall reliability of the facility specific measure "percent of patients with UFR > 13" will achieve a desirable width of confidence interval or margin of error.

The wide average confidence interval indicates that a small number of facilities do not provide enough information to precisely estimate the reliability. Moreover, a small number of facilities leads to biased estimation of the IUR (e.g. the estimated average IUR is smaller than the population based value 0.738). The number of facilities required depends to some degree on the true reliability and the nature of the measure being considered, but this conclusion that a small number of facilities (e.g. less than 30) is insufficient would be quite robust. Of course other considerations besides reliability will also need to enter into measure testing, depending on the context.

In this report, we utilize the accurate estimation of the IUR to determine a minimum number of facilities to ensure a desirable precision. However, it is worth noting that the size of the IUR may not indicate the usefulness of the measure for profiling extreme facilities.[13–15] A measure may have a very low IUR, but could still be useful to identify extreme outcomes. Conversely, a high IUR does not necessarily indicate a wide disparity in the quality of care provided by facilities. In fact, Kalbfleisch et al.[15] show that a large IUR can be a signal of the need for further risk adjustment to account for differences between patients across facilities. Nevertheless, the findings presented in this report have implications in practice. First, the reported results will assist in the planning of future measure testing. Second, we provide methods for computing the IUR for complicated quality measures.

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] Ash SA, Fienberg ES, Louis AT, et al. Statistical Issues in Assessing Hospital Performance: Commissioned by the Committee of Presidents of Statistical Societies. 2011 [Accessed 27 March 2013]. Available from: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf

[2] He K, Kalbfleisch 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] Kalbfleisch 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] 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. 2017; 69(3 Suppl 1): A7-A8. PMid: 28236831. https://doi.org/10.1053/j.ajkd.2016.12.004

[5] Saran R, Bragg-Gresham JL, Levin NW, et al. Longer treatment time and slower ultrafiltration in hemodialysis: associations with reduced mortality in the DOPPS. Kidney International. 2006; 69: 1222-1128. PMid: 16609686. https://doi.org/10.1038/sj.ki.5000186

[6] Movilli E, Gaggia P, Zubani R, et al. Association between high ultrafiltration rates and mortality in uraemic patients on regular

haemodialysis. A 5-year prospective observational multicenter study. Nephrology Dialysis Transplantation. 2007; 22(12): 3547-3552. PMid: 17890254. https://doi.org/10.1093/ndt/gfm466

[7] Flythe JE, Kimmel SE, Brunelli SM. Rapid fluid removal during dialysis is associated with cardiovascular morbidity and mortality. Kidney International. 2011; 79(2): 250-257. PMid: 20927040. https://doi.org/10.1038/ki.2010.383

[8] Burton J, Jefferies HJ, Selby NM. Hemodialysis-Induced Cardiac Injury: Determinants and Associated Outcomes. Clinical Journal of Aamerican Society Nephrology. 2009; 4: 914-920. PMid: 19357245. https://doi.org/10.2215/CJN.03900808

[9] McIntyre CW. Haemodialysis-induced myocardial stunning in chronic kidney disease—a new aspect of cardiovascular disease. Blood Purif. 2010; 29: 105-110. PMid: 20093813. https://doi.org/10.1159/000245634

[10] Adams JL. The reliability of provider profiling: a tutorial. Santa Monica, CA: RAND Corporation; 2009. Available from: http://www.rand.org/pubs/technical_reports/TR653.html

[11] Adams JL, Mehrotra A, Thomas W, et al. Physician cost profiling - reliability and risk of misclassification. N Engl J Med. 2010; 362: 1014-1021. PMid: 20237347. https://doi.org/10.1056/NEJMsa0906323

[12] Yelland LN, Salter AB, Ryan P, et al. Adjusted intraclass correlation coefficients for binary data: methods and estimates from a cluster-randomized trial in primary care. Clin Trials. 2011; 8(1): 48-58. PMid: 21335589. https://doi.org/10.1177/1740774510392256

[13] Staggs VS. Reliability assessment of a hospital quality measure based on rates of adverse outcomes on nursing units. Statistical Methods in Medical Ressearch. 2017; 26(6): 2951-2961. PMid: 26721876. https://doi.org/10.1177/0962280215618688

[14] Staggs VS, Cramer E. Reliability of Pressure Ulcer Rates: How Precisely Can We Differentiate Among Hospital Units, and Does the Standard Signal-Noise Reliability Measure Reflect This Precision? Research in Nursing & Health. 2016; 39(4): 298-305. PMid: 27223598. https://doi.org/10.1002/nur.21727

[15] Kalbfleisch JD, He K, Xia L, et al. Does the inter-unit reliability (IUR) measure reliability? Health Services and Outcomes Research Methodology. 2018; 18(3): 215-225. https://doi.org/10.1007/s10742-018-0185-4