Biometrics 74, 1401–1403
December 2018
DOI: 10.1111/biom.12906

# Discussion on "Time-Dynamic Profiling with Application to Hospital Readmission among Patients on Dialysis," by Jason P. Estes, Danh V. Nguyen, Yanjun Chen, Lorien S. Dalrymple, Connie M. Rhee, Kamyar Kalantar-Zadeh, and Damla Senturk

**John D. Kalbfleisch*** **and Kevin He**

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109-2029, U.S.A.
**email:* jdkalbfl@umich.edu

We congratulate the authors on an interesting paper on an important topic. The importance for quality improvement of monitoring and profiling is great and this article offers a welcome and careful exploration of several aspects of this problem as well as proposing a potentially important time dependent assessment of facility differences. We note, in particular, that the article concentrates on the use of fixed effects methods for estimation of time-dependent facility effects and uses the empirical null as a way to account for extra variation in the Z-scores. These ideas also form the basis of the analysis in He et al. (2013), which is also a key reference to this work.

## 1. Random Effects and Fixed Effects

Hierarchical models with random provider effects are commonly used as the basis of profiling methods. In many instances (e.g., Normand and Shahian, 2007; Ohlssen et al., 2007; Ash et al., 2012), it is proposed that facility effects should be estimated as the mean of the empirical Bayes posterior distribution. We refer to this as the random effects (RE) approach. This leads to shrinkage estimators with their advantage over maximum likelihood based on fixed effects of achieving a smaller mean squared error overall. If, however, one examines the mean squared error conditional on the size of the facility effect, one finds that the overall gain is achieved through gains in the center of the distribution of facility effects, whereas the usual fixed effects estimates are more accurate for the facilities of primary interest, namely those with extreme effects. In addition, naive use of these models can lead to biased estimation of regression coefficients when covariates are correlated with the facility effects. Jones and Spiegelhalter (2011) and Kalbfleisch and Wolfe (2013) discuss these issues in some depth.

In contrast, the fixed effects (FE) approach (e.g., Wolfe et al., 1992; He and Schaubel, 2013) gives more accurate estimates of facility effects for exceptional facilities, and yields unbiased estimates regardless of correlations between facility effects and patient case-mix. The FE approach, however, will tend to identify many more facilities as extreme than one might want since it makes no allowance for unexplained variation between facilities. In fact, both the FE and RE analyses as commonly used entail a test of the sharp null hypothesis that the facility effects conform to a national norm and as a consequence will tend to identify as worse than expected, large facilities, even when their true effect is relatively small. This tends to be a disadvantage of the approach and is one motivation for alternatives such as the empirical null.

## 2. Hospital Discharges and Readmission

A primary reason for monitoring hospital readmissions of dialysis patients is to promote coordination of care between the dialysis facility and the hospital as a patient is discharged following a period of hospitalization (e.g., Wish, 2014). With good coordination, the facility can aid in assuring that treatments recommended at discharge are appropriately followed and that the patient's dialysis treatments will not be disrupted in moving from one provider to the other. To some degree, the rate of readmissions following a discharge can be seen as an indicator of the success of that coordination and post-discharge care. In this view, the reason for hospitalization is relevant and, as a consequence, one would normally condition on patient diagnoses and other current conditions in modeling the chance of readmission. Each time there is a hospital discharge, we begin a new experiment and assess the outcome of that experiment given the conditions at its baseline, that is, at the time of discharge. Estes et al. argue that only covariates that are available at the beginning dialysis should be used in assessing readmission. This view may be better justified for assessing such things as the overall rate of hospitalization or mortality than it is for readmissions, and as noted below, the readmission measure implicitly conditions on the fact that there is a hospital discharge at time $t$.

In fact, another outcome that is monitored for dialysis facilities is the sequence of hospitalizations, or alternatively, the sequence of hospital discharges for patient on dialysis. This is naturally modeled using a counting process $N_{ij}(t)$, which counts the number of hospital discharges in the interval $(0, t)$ for the $j$th patient in the $i$th facility. Here, it could be argued that it is natural to condition primarily on covariates $X_{ij}$ that are measured at the beginning of dialysis; as noted in this article, events that occur over time may be related to the quality

of care administered. A standard proportional rate model could be used for analysis as discussed in Liu et al. (2012) and the aim is to identify facilities whose hospitalization rate is high compared to a national norm.

The model equation (M1) in Estes et al. could be more precisely written by explicitly conditioning on the fact that there is a hospital discharge at time $t$. Thus, the conditioning event is $\{X_{ij}, b_{ij}, dN_{ij}(t) = 1, S_{ij} > t\}$, where $S_{ij}$ is the death time for the $j$th subject in the $i$th facility, and $b_{ij}$ is the subject-specific random effect. One could also define $Y_{ij}(t) = 0$ whenever $dN_{ij}(t) = 0$. In this framework, other covariates available at $t$ could be incorporated into the condition. Both the sequence of hospitalizations and readmissions are useful in assessing dialysis facilities.

## 3. Choice of Time Scale

The addition of the time-dependent facility effect is potentially an important idea. Estes et al. choose the time origin as the beginning of dialysis instead of monitoring the potential variation in facility effects over chronological time. When changes in staffing or management or infection control occur, they are likely to affect all patients in the facility regardless of the time since they began dialysis, so in a way this seems the more natural time scale for assessing variation in facility effects. It seems that it would be possible to revise the analysis to this alternative time scale. One could also entertain models which allow both time scales by adjusting for specific parametric functions of chronological time in the model (M1). Exploration of these two time scales, as discussed for example in Farewell and Cox (1979), could be a useful direction to pursue.

In the example, the patients are all incident and followed for an extended period of time. In most profiling applications, one is more interested in the possibility of quality problems arising over shorter periods of time of perhaps one or two years. In addition, one would normally want to include prevalent patients who are already at risk at the beginning of the observation period in the analysis. Are there any difficulties with extending these methods to this situation? The main difference would seem to be to allow for left truncation to incorporate the prevalent patients.

## 4. Between-Facility Variation and the Empirical Null

A key issue in profiling is the variation in the facility effects, which are denoted by $\gamma_i(t)$ in this article. Although widely used, the term facility effect is a poor one since it carries a connotation of causality that may not be true. In many instances, the majority of the variation in the facility effects is not due to the quality of care but rather to other factors outside the facility's control. For example, there may be differences among the patients treated by different facilities that are not accounted for in the risk adjustment and that affect outcomes; these include such things as genetic makeup or aspects of socioeconomic status or co-morbidities that can vary widely across facilities and are not well measured. Such unexplained variation in facility effects would account for at least some of the over dispersion of the Z-statistics, and unadjusted use of the FE or RE analysis can lead to inappropriately high flagging rates, especially among larger providers. The empirical null provides one way to address this issue.

Insight can be obtained by considering a linear hierarchical model with no covariates, $Y_{ij} = \mu + \alpha_i + \epsilon_i$, where $Y_{ij}$ is the outcome of interest, $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\epsilon_{ij} \sim N(0, \sigma_w^2)$, independently for $i = 1, \ldots, I$ and $j = 1 = \ldots, n_i$. For simplicity, assume that $\sigma_\alpha, \sigma_w$ and $\mu = 0$ are known or estimated with great accuracy as is the case when $I$ is large. Then the mean response in the $i$th facility, $\hat{\alpha}_i = \bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$, is the fixed effects estimate of $\alpha_i$. To account for all the variation among providers, Jones and Spiegelhalter (2011) suggest assessing $\hat{\alpha}_i$ with reference to its marginal distribution $N(0, \sigma_\alpha^2 + \sigma_w^2/n_i)$, including the between-provider variation. We refer to this as the fixed effects analysis with random intercept (FERE). Essentially, this approach allows for the natural variation between providers and does not flag a provider unless it is extreme with reference to the total variation. This substantially reduces the tendency to flag large providers with moderate values of $\alpha_i$ and the bias against large providers in the FE and the RE approaches is reduced. If $n_i = n$, for all $i = 1, \ldots, I$, then it can be seen that the empirical null approach would yield asymptotically the same result as FERE, so that it also accounts for the total variation. This is similar to arguments in Efron (2004). As Estes et al. note, the empirical null distribution depends on the sample size and the approach can be implemented within strata in order to obtain relatively homogeneous groups of facilities. See also, Kalbfleisch and Wolfe (2013) and He et al. (2013) for additional discussion on this point.

We make a few comments regarding the empirical null approach:

1. In practice, we encounter various types of patient outcomes, and the FERE approach may be difficult to implement. The empirical null approach generalizes relatively easily as is illustrated in this article.
2. The empirical null approach is more robust against outliers. The existence of outliers will result in an overestimate of the between-facility variance in the normal case discussed above, but will affect much less the robust estimation used in the empirical null.
3. Stratification on sample size is not a fully satisfactory approach since the number of bins selected is arbitrary and will affect flagging rules. A related problem is the discontinuity of the critical line at the boundaries between the strata where, for example, two providers near a boundary may have similar Z-scores and sample sizes, but different flagging status. To overcome these issues, with our colleagues Lu Xia and Yanming Li, we have been developing smoothed estimates of the mean and variance of the Z-scores as a function of sample size so that each provider has an individualized empirical null distribution.
4. For profiling, both FERE and the empirical null techniques allow for all of the between-facility variation. On the other hand, the FE and RE approaches essentially assume that all the between-facility variation is due to quality of care and allows for no unexplained or natural variation. Kalbfleisch et al. (2018) consider intermediate approaches that specify some proportion of the

between-facility variation as being due to the quality of care provided.

## 5. Some Issues with Standardized Measures

In developing readmission measures for dialysis facilities, an important consideration is that both hospitals and dialysis facilities play a key role in averting unnecessary readmissions. Further, patients from a dialysis facility may be admitted to various hospitals and similarly, hospitals may receive patients from multiple dialysis facilities. In addition, hospitals vary in their readmission rates as documented in the Hospital Compare measure of the Centers for Medicare and Medicaid Services (CMS) (see, Horwitz et al., 2011). In order to account for this, He et al. (2013) included hospital as a random effect and found that this resulted in a substantial improvement in fit. It should be noted that several previous studies have suggested an influence of multiple types of providers on treatment practices and outcomes among ESRD patients (Hirth et al., 2009; Hirth et al., 2010; Turenne et al., 2010). Estes et al. includes a random effect for patients, which seems particularly appropriate and important when patients are followed over a long period of time. He et al. (2013) examine readmission on prevalent patients over a shorter period of time where random effects for patients seemed less important. Computational issues in fitting models with random effects for both patients and hospitals and potentially for another providers would be of substantial interest.

Estes et al. caution that it is generally wrong to compare indirect standardized measures between two facilities, because each facility's indirect standardized measure is essentially adjusted to a different (facility-specific) covariate distribution. In effect, we can define a standardized ratio measure with respect to any distribution of the covariates that we wish, so that we could define the SDRR for facility $i$ with reference to the patient population in facility $i'$. This would be estimated with

$$\widehat{SDRR}_i(t;i') = \frac{\sum_{j\in N_{i't}} g^{-1}(\hat{\gamma}_i(t) + \hat{b}_{i'j} + Z_{i'j}^T\hat{\beta})}{\sum_{j\in N_{i't}} g^{-1}(\hat{\gamma}_M(t) + \hat{b}_{i'j} + Z_{i'j}^T\hat{\beta})}.$$

This quantity would be directly comparable to $\widehat{SDRR}_{i'}(t)$. As noted in Estes et al., the SDRR of facility $i$ will not change much if the covariate distribution in $i'$ is similar to that in $i$, but this gives a more formal check. Note that one could also argue that the comparison of the SDRR at time $t$ to that at a different time $t'$ is also subject to the same change in covariate distribution and the same cautions. An alternative approach is to adjust all measures to the overall distribution of covariates combining across all facilities. This makes all measures comparable and corresponds to direct standardization. The potential difficulties with comparisons across facilities notwithstanding, we prefer indirect standardization as giving a more meaningful measure to each facility comparing their results with the national norm for the patients they actually treat.

Finally, we want to thank these authors for their work on this important problem and for considering some generalizations of standard profiling approaches that are potentially important in applications.

## References


Farewell, V. T., and D. R. Cox. (1979). A Note on Multiple Time Scales in Life Testing. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **28**, 73–75.

Hirth, R. A., Turenne, M. N., Wheeler, J., Pan, Q., Ma, Y., and Messana, J. M. (2009). Provider monitoring and pay-for-performance when multiple providers affect outcomes: An application to renal dialysis. *Health Services Research* **44**, 1585–1602.

Hirth, R. A., Turenne, M. N., Wheeler, J., Ma, Y., and Messana, J. M. (2010). Do resource utilization and clinical measures still vary across dialysis chains after controlling for the local practices of facilities and physicians? *Medical Care* **48**, 726–732.

Jones, H. E., and Spiegelhalter, D. J. (2011). The identification of unusual health-care providers from a hierarchical model. *The American Statistician* **65**, 154–163.

Kalbfleisch, J. D., He, K., Xia, L., and Li, Y. (2018). Does the inter-unit reliability (IUR) measure reliability? Submitted.

Liu, D., Schaubel D. E., and Kalbfleisch, J. D. (2012). Computationally efficient marginal models for clustered recurrent event data. *Biometrics* **68**, 637–647.

Ohlssen, D. I., Sharples, L. D., and Spiegelhalter, D. J. (2007). A hierarchical modelling frame work for identifying unusual performance in health care providers. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **170**, 865–890.

Turenne, M. N., Hirth, R. A., Messana, J. M., Turner, J. S., Sleeman, K. K., and Wheeler, J. (2010). When payment systems collide: The effect of hospitalization on anemia in renal dialysis patients. *Medical Care* **48**, 296–305.

Wish, J. G. (2014). The role of thirty day readmission as a measure of quality. *Clinical Journal of the American Society of Nephrology* **9**, 440–442.

Wolfe, R. A., Gaylin, D. S., Port, F. K., Held, P. J., and Wood, C. L. (1992). Using USRDS generated mortality tables to compare local ESRD mortality rates to national rates. *Kidney International* **42**, 991–996.