HSR Health Services Research

DOI: 10.1111/1475-6773.12726
RESEARCH BRIEF

A Five-Tier System for Improving the Categorization of Transplant Program Performance

Andrew Wey, Nicholas Salkowski, Bertram L. Kasiske, Ajay K. Israni, and Jon J. Snyder

Objective. To better inform health care consumers by better identifying differences in transplant program performance.

Data Source. Adult kidney transplants performed in the United States, January 1, 2012–June 30, 2014.

Study Design. In December 2016, the Scientific Registry of Transplant Recipients instituted a five-tier system for reporting transplant program performance. We compare the differentiation of program performance and the simulated misclassification rate of the five-tier system with the previous three-tier system based on the 95 percent credible interval.

Data Collection. Scientific Registry of Transplant Recipients database.

Principal Findings. The five-tier system improved differentiation and maintained a low misclassification rate of less than 22 percent for programs differing by two tiers.

Conclusion. The five-tier system will better inform health care consumers of transplant program performance.

Key Words. Rating system, transplant program performance

There is considerable interest in providing health care consumers with "report cards" on the performance of solid-organ transplant programs (Smits et al. 2003; Snyder et al. 2016a, b) and other medical providers (Racz and Sedransk 2010). The Scientific Registry of Transplant Recipients (SRTR) publishes program-specific reports on the performance of solid-organ transplant programs every 6 months. Historically, SRTR has emphasized summarizing program performance with statistical summary measures, for example, hazard ratios and 95 percent credible intervals. However, the Agency for Healthcare Research and Quality (AHRQ) recommends that information targeting health

care consumers should avoid statistical summary measures (Hibbard and Sofaer 2010). Thus, SRTR recently instituted a public reporting system that places programs into one of five easily interpretable tiers with similar levels of over- or underperformance within each tier.

SRTR previously categorized transplant programs into three tiers that corresponded to "lower than expected," "as expected," or "higher than expected" survival. The categorization was based on whether the 95 percent credible interval of the hazard ratio included 1. This approach led to strict definitions of over- and underperformance for placing programs into tier 1 or 3, which strongly depended on the number of transplants performed by the program. These strict definitions prevented adequate differentiation of program performance, as tier 2 included point estimates of the hazard ratios ranging from 0.5 to 2. Thus, the three-tier system may not appropriately inform health care consumers regarding the performance of transplant programs or other medical providers.

SRTR instituted a five-tier system to better differentiate program performance by increasing the number of tiers and expanding the interval definitions. The five-tier system will lead to better public reporting of transplant program performance than the previous three-tier system by better identifying programs with similar performance. In this study, we demonstrate that the five-tier system improved differentiation of transplant program performance and maintained a relatively low misclassification rate compared with the previous three-tier system based on the 95 percent credible interval.

Address correspondence to Andrew Wey, Ph.D., Scientific Registry of Transplant Recipients, Minneapolis Medical Research Foundation, 701 Park Avenue, Suite S4.100, Minneapolis, MN 55415. e-mail: awey@cdrg.org. Nicholas Salkowski, Ph.D., is with the Scientific Registry of Transplant Recipients, Minneapolis Medical Research Foundation, Minneapolis, MN. Bertram L. Kasiske, M.D., is with the Department of Medicine, Hennepin County Medical Center, Minneapolis, MN, and Scientific Registry of Transplant Recipients, Minneapolis Medical Research Foundation, Minneapolis, MN. Ajay K. Israni, M.D., M.S., is with the Scientific Registry of Transplant Recipients, Minneapolis Medical Research Foundation, Minneapolis, MN; Department of Medicine, Hennepin County Medical Center, Minneapolis, MN, and Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN. Jon J. Snyder, Ph.D., is with the Scientific Registry of Transplant Recipients, Minneapolis Medical Research Foundation, Minneapolis, MN, and Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN.

METHODS

This study used data from the SRTR. The SRTR data system includes data on all donors, waitlisted candidates, and transplant recipients in the United States, submitted by the members of the Organ Procurement and Transplantation Network (OPTN), and has been described elsewhere (Leppke et al. 2013). The Health Resources and Services Administration, US Department of Health and Human Services, provides oversight of the activities of the OPTN and SRTR contractors.

We used the Bayesian methodology currently to evaluate transplant program performance (Salkowski et al. 2014a, b). Transplant programs are evaluated on the number of observed events versus the number of expected events as estimated by a risk-adjustment model (Snyder et al. 2016a, b). The number of observed events follows a Poisson distribution, which motivates a conjugate Gamma prior partly to simplify estimation of program-specific posterior distributions. Thus, the posterior distribution for program-specific hazard ratios is a Gamma distribution that depends only on the number of observed and expected events (Salkowski et al. 2014a, b).

Categorizing Programs Based on Performance

Programs are categorized into tiers of relative performance based on estimated program-specific ratings ($R_i^{(k)}$), where k is the number of tiers. The ratings are estimated by taking the expectation of a monotonically decreasing function with respect to the posterior distribution of the program's hazard ratio. The five- and three-tier systems used a logistic-type utility function and an indicator function, respectively (see Supporting Information for additional technical details). Programs are then placed into tiers of relative performance with higher tiers suggesting better post-transplant survival. The three- and five-tier systems were relatively symmetric; for example, the level of underperformance for programs in the second tier was similar to the level of overperformance for programs in the fourth tier.

As the ratings are between 0 and 1, the three-tier system based on the 95 percent credible interval is

$$\textit{Three-Tier System} = \begin{cases} \textit{Tier}\,1\,(\textit{Worse than Expected}), \text{if } R_i^{(3)} < 0.025 \\ \textit{Tier}\,2\,(\textit{As Expected}), \text{if } R_i^{(3)} \geq 0.025 \text{ and } R_i^{(3)} < 0.975 \\ \textit{Tier}\,3\,(\textit{Better than Expected}), \text{if } R_i^{(3)} \geq 0.975. \end{cases}$$

The three-tier system placed the vast majority of transplant programs into tier 2, or "performing as expected," due to the strict cutoffs for tiers 1 and 3. This caused significant variability in the performance of tier 2 programs and, ultimately, poor differentiation of the three-tier system. Another drawback of the three-tier system was that small programs were unlikely to be in a better- or worse-than-expected tier regardless of observed program outcomes. Small programs are inherently more likely than large programs to be classified as "performing as expected" due to less information in the posterior. However, the conservativeness of the three-tier system could mislead consumers regarding the performance of small programs.

Classification of program performance is ultimately a trade-off between maximizing the differentiation of program performance and minimizing the misclassification, or incorrect ranking, of programs. Specifically, a larger number of categories will generally improve differentiation of program performance at the cost of higher misclassification. Yet a larger number of categories are generally more difficult for consumers to comprehend (Hibbard and Sofaer 2010). Thus, to improve differentiation of program performance while maintaining accessible public reporting, SRTR chose a five-tier system for categorizing program performance. A five-tier system allows SRTR to identify over- and underperformance in small- to moderate-sized programs, while distinguishing between programs with exceptional over- or underperformance. Specifically, the five-tier system is:

$$\textit{Five-Tier System} = \begin{cases} \textit{Tier}\,1\,(\textit{Worse than Expected}), \text{if } R_i^{(5)} < 0.125 \\ \textit{Tier}\,2\,(\textit{Somewhat worse than Expected}), \text{if } R_i^{(5)} \geq 0.125 \text{ and } R_i^{(5)} < 0.375 \\ \textit{Tier}\,3\,(\textit{As Expected}), \text{if } R_i^{(5)} \geq 0.375 \text{ and } R_i^{(5)} < 0.625 \\ \textit{Tier}\,4\,(\textit{Some what Better than Expected}), \text{if } R_i^{(5)} \geq 0.625 \text{ and } R_i^{(5)} < 0.875 \\ \textit{Tier}\,5\,(\textit{Better than Expected}), \text{if } R_i^{(5)} \geq 0.875. \end{cases}$$

Evaluation of Kidney Transplant Programs

The five-tier system was illustrated with 1-year post-transplant graft survival for adult kidney transplant programs operating between January 1, 2012, and June 30, 2014, with data from the June 2015 SRTR standard analytic file (SAF).

We compared categorization of program performance in the five-tier system with categorization in the previous three-tier system. The distribution of programs and hazard ratios in each tier was investigated across program volume for the five- and three-tier systems. The within-tier sum-of-squares estimated the differentiation of program-specific hazard ratios for both systems. The within-tier sum-of-squares is the within-cluster sum-of-squares in which the tiers are the clusters. The within-cluster sum-of-squares selects the optimal clusters in the *k*-means clustering algorithm (Everitt and Dunn 2001). Thus, lower values for the within-tier sum-of-squares suggest better differentiation of hazard ratios.

Probability of Accurate Categorization

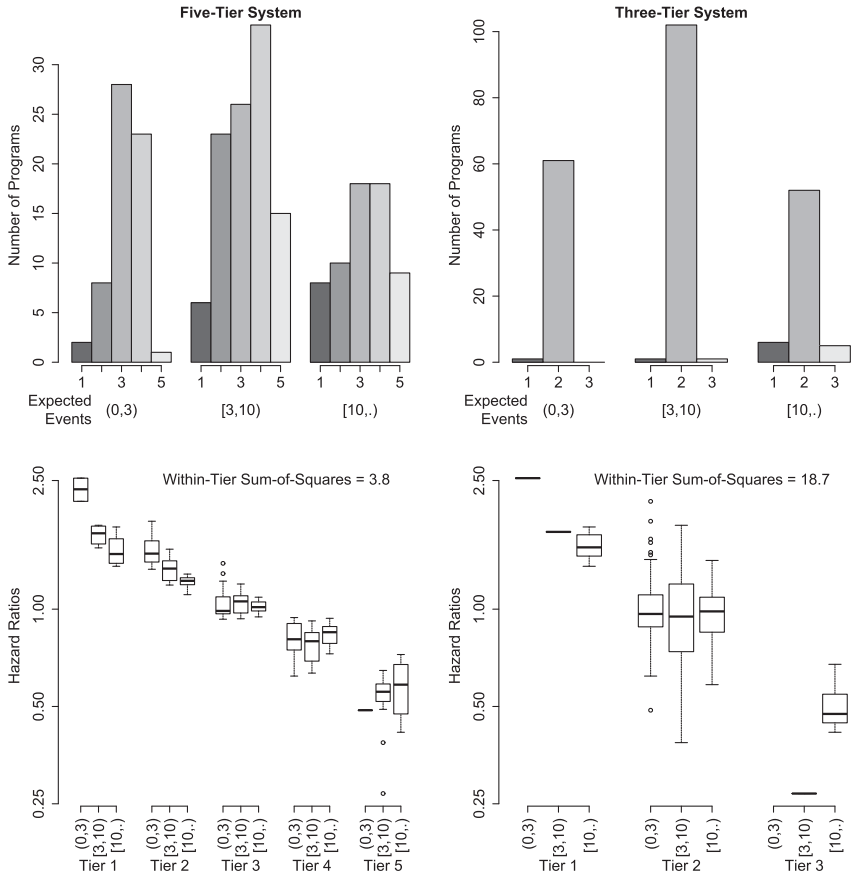
A simulation study evaluated the ability of the five- and three-tier systems to accurately categorize programs. Specifically, categorization of adult kidney transplant programs was simulated for transplants performed between January 1, 2012, and June 30, 2014, by randomly generating sets of 229 programs. The number of expected events was randomly sampled from the number of expected events from adult kidney programs. The hazard ratios for each program were randomly selected from a Gamma distribution with mean 1 and variance 1/8, which approximates the mean and variance of the observed hazard ratios for adult kidney programs. The observed number of events was simulated from a Poisson distribution with a mean equal to the number of expected events times the hazard ratio of the program. We then estimated the probability that a given tier program had a better hazard ratio than a lower-tier program. Thus, the simulation estimated, for example, the probability that a tier 4 program was truly better than a tier 2 program. The simulation was completed 2,000 times to minimize the effect of randomly sampling hazard ratios, program sizes, and observed events. The Supporting Information presents a sensitivity analysis for different levels of variability and underlying hazard ratio distributions.

RESULTS

Evaluation of Kidney Transplant Programs

Figure 1 (top panels) presents the distribution of tiers for the five-tier and three-tier systems. The three-tier system categorized eight and six programs into tiers 1 and 3, respectively, and most of these programs experienced

Figure 1: The Top Panels Present the Number of Programs Placed in Each Tier for the Instituted Five-Tier System (Left Panels) and the Previous Three-Tier System (Right Panels)



Notes. The bottom panels illustrate the distribution of hazard ratios within each tier while stratified by the number of expected events. The bottom panels present the within-tier sum-of-squares for both systems.

more than 10 expected events. In contrast, the five-tier system categorized 16 and 25 programs into tiers 1 and 5, respectively, and small-volume programs were still less likely to be in tier 1 or 5. Overall, the five-tier system was less likely to place a program in the "performing as expected" tier regardless of volume.

Figure 1 (bottom panels) presents the distribution of hazard ratios for the five- and three-tier systems stratified by number of expected events. The five-tier system improved the differentiation of program-specific hazard ratios compared with the three-tier system. For example, the hazard ratios were relatively well ordered by the five-tier system with higher hazard ratios in lower tiers. In contrast, the three-tier system showed larger variability in the hazard ratios of tier 2 programs, ranging from 0.5 to 2.0. The five-tier system reduced the within-tier sum-of-squares by 80 percent compared with the three-tier system. Thus, programs in the same tier of the five-tier system have relatively similar performance, especially compared with programs in the same tier of the three-tier system, and the five-tier system differentiates observed hazard ratios better than the three-tier system.

Probability of Accurate Categorization

Table 1 presents the probability of accurately categorizing programs. For example, in the three-tier system, a tier 3 program has a probability of 88 percent and 100 percent of having a better hazard ratio than a tier 2 and tier 1 program, respectively. In contrast, in the five-tier system, a tier 5 program has a 91 percent and 98 percent probability of having a better hazard ratio than a tier 2 and tier 1 program, respectively. However, one-tier differences have relatively low probabilities of correct categorization, ranging from a 63 percent probability that a tier 4 program is better than a tier 3 program to a 75 percent probability that a tier 2 program is better than a tier 1 program. Two-tier differences have higher probabilities of correct categorization, ranging from a 78 percent probability that a tier 4 program is better than a tier 2 program to an

Table 1: The Probability That a Program in the Given Tier of the Column Is Truly Better Than a Program in the Given Tier of the Row

System	*Tier*	*5*	*4*	*3*	*2*
Five-tier	4	0.72	–	–	–
	3	0.81	0.63	–	–
	2	0.91	0.78	0.66	–
	1	0.98	0.92	0.86	0.75
Three-tier	2	–	–	0.85	–
	1	–	–	1.00	0.90

Note: The program with a lower hazard ratio is the better program. The probability is presented for both the instituted five-tier system and previous three-tier system.

86 percent probability that a tier 3 program is better than a tier 1 program. The five-tier system therefore improved the differentiation of program evaluations while maintaining relatively good classification, especially for two- to four-tier differences.

Additional simulation scenarios in the Supplementary Materials illustrate the performance of the five- and three-tier systems under varying levels of hazard ratio variability and alternative distributional assumptions. Both the five- and three-tier systems better classified program performance with more variability in the underlying hazard ratios, while the results were similar under alternative distributional assumptions. The Supplementary Materials present an additional simulation study that compares the classification of small and large programs between the three- and five-tier systems. The probability that a large program was placed in a higher tier than a small program increased with worse performance of the small program for the five-tier system but not for the three-tier system.

DISCUSSION

In December 2016, SRTR instituted a five-tier system to categorize transplant program performance and to better identify meaningful differences among providers. Compared with the three-tier system, the five-tier system improves differentiation of program performance by simultaneously adding additional tiers and expanding interval definitions to include more programs. The improved differentiation ensures that meaningful differences between providers are adequately identified and not obscured by an inability to identify over- or underperformance in small- to moderate-volume programs. The Supporting Information demonstrates that the five-tier system reduces the within-tier sum-of-squares for the other major solid-organ transplants, that is, liver, heart, and lung transplants. The five-tier system will therefore provide a better basis for health care consumers to make informed decisions regarding solid-organ transplantation by improving the accessibility and clinical interpretation of the public reporting.

Any system that categorizes programs based on relative performance faces a trade-off between maximizing the differentiation of program performance and minimizing the misclassification of programs for over- or underperformance. For example, a system with more tiers will usually show better differentiation but higher misclassification than a system with fewer tiers. The number of tiers that optimally balances differentiation and misclassification

may be estimable by, for example, extending the "gap statistic" to program differentiation (Tibshirani, Walther, and Hastie 2001). However, there is evidence that fewer categories can improve consumer understanding of report cards (Hibbard and Sofaer 2010). Thus, the determination of the optimal number of tiers should consider technical and practical constraints and remains an open research question.

The evaluation of transplant program performance has potential unintended consequences (Hamilton 2013; Schold et al. 2013, 2016; Reese et al. 2016). These consequences fundamentally arise from the lower rate of graft and patient survival for high-risk donors and recipients. The transplant community perceives that high-risk donors and recipients place programs at risk for regulatory review despite adjustments for donor and recipient risk factors (Reese et al. 2016; Snyder et al. 2016a, b), which should mitigate the concerns and allow programs to perform transplants with higher measured risk. The five-tier system may increase concerns about unintended consequences due to improved differentiation of program performance. However, unless a program truly underperforms, every program has an incentive to perform more transplants to improve the precision of the posterior and the corresponding certainty that the program is performing as expected or better. Thus, the unintended consequences of evaluating program performance can be minimized through better education regarding risk adjustment and the importance of precision in evaluating programs with truly average or better outcomes.

There are general limitations to placing transplant programs and other health care providers into tiers of relative performance. First, adequate risk adjustment for donor and recipient characteristics is critical in ensuring fair comparison and categorization. The process of building risk-adjustment models in transplantation is thorough, with consideration of numerous potential risk factors and nonlinear terms for continuous factors (Snyder et al. 2016a, b). However, important risk factors may be poorly measured or not collected. It is therefore critical that accurate and complete data are reported on important risk factors to ensure appropriate risk-adjusted evaluations. Second, the loss of information due to transforming the posterior distribution into a rating and then placing each rating into a tier does not guarantee that a higher ranking implies probabilistically better performance. This trade-off is necessary to ensure the accessibility of transplant program evaluations for health care consumers. Lastly, the Gamma–Poisson model for estimation of the posterior distribution of program-specific hazard ratios could be misspecified due to, for example, overdispersion. Severe misspecification could affect the simulated

misclassification rates, which assumed program hazard ratios were Gamma-distributed.

The three-tier system based on the 95 percent credible interval is implemented in other medical fields, such as cardiac surgery (Racz and Sedransk 2010; The Society of Thoracic Surgeons, 2016). However, unless every medical provider performs many operations, the three-tier system will have poor differentiation of provider performance. As a system that fails to identify meaningful differences in program performance will fail to adequately inform health care consumers, a five-tier system could improve assessment of health care providers outside the field of transplantation. Each medical field should determine the appropriate balance between differentiation of provider performance and the risk of misclassification. This may help inform health care consumers regarding provider performance in transplantation and other medical fields.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work was conducted under the auspices of the Minneapolis Medical Research Foundation, contractor for the Scientific Registry of Transplant Recipients, as a deliverable under contract number HHSH250201500009C (US Department of Health and Human Services, Health Resources and Services Administration, Healthcare Systems Bureau, Division of Transplantation). As a U.S. Government–sponsored work, there are no restrictions on its use. The views expressed herein are those of the authors and not necessarily those of the U.S. Government. AKI was partially supported by R01 HS 24527.

Disclosures: None.

Disclaimer: None.

REFERENCES

Everitt, B. S., and G. Dunn. 2001. *Applied Multivariate Data Analysis*, 2nd Edition. Chichester, West Sussex, UK: John Wiley and Sons.

Hamilton, T. E. 2013. "Regulatory Oversight in Transplantation: Are the Patients Really Better off?" *Current Opinion in Organ Transplantation* 18 (2): 203–9.

Hibbard, J., and S. Sofaer. 2010. "Best Practices in Public Reporting no. 1: How to Effectively Present Health Care Performance Data to Consumers" [accessed on

January 5, 2017]. Available at http://archive.ahrq.gov/professionals/quality-patient-safety/quality-resources/tools/pubrptguide1/pubrptguide1.pdf

Leppke, S., T. Leighton, D. Zaun, S. C. Chen, M. Skeans, A. K. Israni, and B. L. Kasiske. 2013. "Scientific Registry of Transplant Recipients: Collecting, Analyzing, and Reporting Data on Transplantation in the United States." *Transplantation Reviews* 27 (2): 50–6.

Racz, M. J., and J. Sedransk. 2010. "Bayesian and Frequentist Methods for Provider Profiling Using Risk-Adjusted Assessments of Medical Outcomes." *Journal of the American Statistical Association* 105 (489): 48–58.

Reese, P. P., M. N. Harhay, P. L. Abt, H. L. Matthew, and S. D. Halpern. 2016. New Solutions to Reduce Discard of Kidneys Donated for Transplantation. *Journal of the American Society of Nephrology* 27 (4): 973–80.

Salkowski, N., J. J. Snyder, D. A. Zaun, T. Leighton, E. B. Edwards, A. K. Israni, and B. L. Kasiske. 2014a. "A Scientific Registry of Transplant Recipients Bayesian Method for Identifying Underperforming Transplant Programs." *American Journal of Transplantation* 14 (6): 1310–7.

Salkowski, N., J. J. Snyder, D. A. Zaun, T. Leighton, A. K. Israni, and B. L. Kasiske. 2014b. Bayesian Methods for Assessing Transplant Program Performance. *American Journal of Transplantation* 14 (6): 1271–6.

Schold, J. D., L. D. Buccini, E. D. Poggio, S. M. Flechner, and D. A. Goldfarb. 2016. "Association of Candidate Removals from the Kidney Transplant Waiting List and Center Performance Oversight." *American Journal of Transplantation* 16 (4): 1276–84.

Schold, J. D., L. D. Buccini, T. R. Srinivas, R. T. Srinivas, E. D. Poggio, S. M. Flechner, C. Soria, D. L. Segev, J. Fung, and D. A. Goldfarb. 2013. "The Association of Center Performance Evaluations Kidney Transplant Volume in the United States." *American Journal of Transplantation* 13 (1): 67–75.

Smits, J., J. De Meester, M. Deng, H. Scheld, M. Hummel, F. Schoendube, A. Haverich, J. Vanhaecke, and H. C. vanHouwelingen; COCPIT Study Group; Eurotransplant heart transplant programs. 2003. Mortality Rates after Heart Transplant: How to Compare Center-Specific Outcome Data? *Transplantation.* 75 (1): 90–6.

Snyder, J. J., N. Salkowski, S. J. Kim, D. Zaun, H. Xiong, A. K. Israni, and B. L. Kasiske. 2016a. "Developing Statistical Models to Assess Transplant Outcomes Using National Registries: the Process in the United States." *Transplantation* 100 (2): 288–94.

Snyder, J. J., N. Salkowski, A. Wey, A. K. Israni, J. D. Schold, D. L. Segev, and B. L. Kasiske. 2016b. "Effects of High-Risk Kidneys on Scientific Registry of Transplant Recipients Program Quality Reports." *American Journal of Transplantation* 16 (9): 2646–53.

The Society of Thoracic Surgeons. 2016. "Explanation of STS CHSD Star Ratings" [accessed on January 5, 2017]. Available at http://www.sts.org/quality-research-patient-safety/sts-public-reporting-online/explanation-sts-chsd-star-ratings

Tibshirani, R., G. Walther, and T. Hastie. 2001. "Estimating the Number of Clusters in a Data Set Via the Gap Statistic." *Journal of the Royal Statistical Society, Series B* 63 (2): 411–23.

University of Michigan Kidney Epidemiology and Cost Center. 2015. "Technical Notes on Dialysis Facility Compare Star Rating System" [accessed on January 5, 2017]. Available at https://dialysisdata.org/sites/default/files/content/Methodology/StarRatings.pdf

SUPPORTING INFORMATION

Additional supporting information may be found online in the supporting information tab for this article:

Appendix SA1: Author Matrix.

Methods S1. Estimation of Program Ratings.

Methods S2. Additional Simulation: Comparing Programs of Different Volume.

Methods S3. Sensitivity Analysis.

Methods S4. Implementation for Liver, Heart, and Lung Transplant Programs.

Figure S1. For the Evaluation of Liver Transplant Programs, the Top Panels Present the Number of Programs Placed in Each Tier for the Instituted Five-Tier System (Left Panels) and the Previous Three-Tier System (Right Panels).

Figure S2. For the Evaluation of Heart Transplant Programs, the Top Panels Present the Number of Programs Placed in Each Tier for the Instituted Five-Tier System (Left Panels) and the Previous Three-Tier System (Right Panels).

Figure S3. For the Evaluation of Lung Transplant Programs, the Top Panels Present the Number of Programs Placed in Each Tier for the Instituted Five-Tier System (Left Panels) and the Previous Three-Tier System (Right Panels).

Table S1. The Probability That a Program with the Given Number of Expected Events (Ten or Forty) has a Better Rating Than a Small Program (Three Expected Events) for the Given Hazard Ratio Combinations. The Hazard Ratios in the Column Correspond to the Small Program with Three Expected Events.

Table S2. A Sensitivity Analysis for the Probability That a Program in the Given Tier of the Column Is Truly Better Than a Program in the Given Tier of the Row.

Table S3. A Sensitivity Analysis for the Probability That a Program in

the Given Tier of the Column Is Truly Better Than a Program in the Given Tier of the Row.

Table S4. A Sensitivity Analysis for the Probability That a Program in the Given Tier of the Column Is Truly Better Than a Program in the Given Tier of the Row.

Table S5. A Sensitivity Analysis for the Probability That a Program in the Given Tier of the Column Is Truly Better Than a Program in the Given Tier of the Row.