American Journal of Transplantation 2014; 14: 1310–1317
Wiley Periodicals Inc.

doi: 10.1111/ajt.12702

A Scientific Registry of Transplant Recipients Bayesian Method for Identifying Underperforming Transplant Programs

N. Salkowski[1], J. J. Snyder[1,2], D. A. Zaun[1], T. Leighton[1], E. B. Edwards[3], A. K. Israni[1,2,4] and B. L. Kasiske[1,4,*]

[1] Scientific Registry of Transplant Recipients, Minneapolis Medical Research Foundation, Minneapolis, MN
[2] Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN
[3] Organ Procurement and Transplantation Network, United Network for Organ Sharing, Richmond, VA
[4] Department of Medicine, Hennepin County Medical Center, Minneapolis, MN
[*] Corresponding author: Bertram L. Kasiske, kasis001@umn.edu

In response to recommendations from a recent consensus conference and from the Committee of Presidents of Statistical Societies, the Scientific Registry of Transplant Recipients explored the use of Bayesian hierarchical, mixed-effects models in assessing transplant program performance in the United States. Identification of underperforming centers based on 1-year patient and graft survival using a Bayesian approach was compared with current observed-to-expected methods. Fewer small-volume programs (<10 transplants per 2.5-year period) were identified as underperforming with the Bayesian method than with the current method, and more mid-volume programs (10–249 transplants per 2.5-year period) were identified. Simulation studies identified optimal Bayesian-based flagging thresholds that maximize true positives while holding false positive flagging rates to approximately 5% regardless of program volume. Compared against previous program surveillance actions from the Organ Procurement and Transplantation Network Membership and Professional Standards Committee, the Bayesian method would have reduced the number of false positive program identifications by 50% for kidney, 35% for liver, 43% for heart and 57% for lung programs, while preserving true positives for, respectively, 96%, 71%, 58% and 83% of programs identified by the current method. We conclude that Bayesian methods to identify underperformance improve identification of programs that need review while minimizing false flags.

Keywords: Graft survival, quality assurance, solid organ transplantation

Abbreviations: CMS, Centers for Medicare & Medicaid Services; E, expected; HR, hazard ratio; MPSC, Membership and Professional Standards Committee; O, observed; OPTN, Organ Procurement and Transplantation Network; PAIS, Performance Analysis and Improvement Subcommittee; PSR, program-specific reports; SRTR, Scientific Registry of Transplant Recipients; UNOS, United Network for Organ Sharing

Received 07 August 2013, revised 09 December 2013 and accepted for publication 30 December 2013

Introduction

Solid organ transplantation is the treatment of choice for most suitable candidates with end-stage organ disease. However, there is a shortage of donor organs, and many candidates die without undergoing transplant. The US Government is charged with equitably distributing donor organs and ensuring optimal survival of patients and organs from both deceased and living donors. It is essential for potential donors, donor families, patients and the general public to know that the best use is made of this scarce resource.

The National Organ Transplantation Act (1984 Pub.L. 98–507) created the Organ Procurement and Transplantation Network (OPTN) and the Scientific Registry of Transplant Recipients (SRTR) (1). OPTN is charged with optimizing deceased donor organ allocation and developing policies to ensure the best possible outcomes for solid organ transplants. SRTR provides data analyses necessary for the Secretary of Health and Human Services to conduct ''an ongoing evaluation of the scientific and clinical status of organ transplantation'' (42 USC 274a). As part of this charge, SRTR is required to produce semiannual reports of transplant program performance (42 USC §121.11(b)). These reports include information on risk-adjusted graft and patient survival and ''confidence intervals or other measures that provide information on the extent to which chance may influence transplant program-specific results'' (42 USC §121.11(b)).

SRTR identifies potentially underperforming programs based on analyses of 1-year patient and graft survival rates (2). Every 6 months, SRTR calculates expected patient deaths and graft failures for each transplant program and compares these to observed patient deaths and graft failures. This comparison is used to identify, or ''flag,'' programs that may be underperforming so the OPTN Membership and Professional Standards Committee (MPSC) can review them further. In addition, these program-specific reports (PSR) are used by the Centers for Medicare & Medicaid Services (CMS), private insurance providers and the general public to assess transplant program outcomes.

SRTR and OPTN hosted a consensus conference on program quality and surveillance February 13–15, 2012 (3). One of the key recommendations of this conference was to explore the use of Bayesian hierarchical, mixed-effects statistical methods to assess program performance. Coincidentally, a report commissioned by CMS to the Committee of Presidents of Statistical Societies, published in January 2012, also recommended using Bayesian hierarchical, mixed-effects models in assessing hospital performance (4). Therefore, SRTR examined the use of these statistical methods in its PSR.

Identifying Underperforming Programs With Flagging Thresholds

Regulatory bodies ask a deceptively simple question: ''Which programs should be reviewed due to outcomes that appear to be significantly worse than expected?'' The answer depends, in part, on how *expected* performance is determined, and that determination depends on the statistical methods employed and on the set of variables used for risk adjustment. Also, the word *significantly* could refer to clinical or to statistical significance. Clinical significance denotes a level of performance that is worse than expected to a clinically meaningful extent. Statistical significance means that a program's observed underperformance is unlikely to be the result of chance alone.

Current method for identifying underperforming transplant programs

SRTR provides risk-adjusted transplant outcomes to help the MPSC determine which programs to review further. These methods have previously been described in detail (2). Currently, a program that performs at least 10 transplants during the 2.5-year evaluation period will be flagged if all of the following three criteria are met:

1. The number of observed (O) outcomes (graft failures or deaths) is more than three more than expected (E) $(O - E > 3)$.
2. The O is more than 50% higher than E $(O/E > 1.5)$.
3. The probability that this observation occurred by random chance is less than 5% (one-sided hypothesis test for $O \leq E$ has a p-value < 0.05).

A program that performs fewer than 10 transplants over a 2.5-year period will be flagged if at least one event occurs. The current system used by the MPSC considers small-volume programs (<10 transplants over a 2.5-year period) separately because the standard methods will likely not identify small underperforming programs, given the difficulty in meeting statistical significance thresholds with small sample sizes. Therefore, SRTR flags small-volume programs that experience at least one event within 1 year of transplant, and the MPSC further scrutinizes these programs if one additional event occurs in the next 2.5-year observation period (advancing by 6 months).

Bayesian method for identifying underperforming transplant programs

The current method produces a yes-or-no decision to identify a program possibly needing further evaluation, while the Bayesian method produces a bell-shaped curve indicating the likely performance of a program relative to the national standard. Regulatory authorities must decide, based on the shape and location of this bell-shaped curve, whether they believe a program requires further scrutiny. Christiansen and Morris provide an overview of Bayesian methods used to design clinically meaningful triggers to further assess health-care provider performance (5). In general, we want to choose a threshold (or thresholds) that, if exceeded, would trigger review. For example, in our field we are interested in a transplant program's performance relative to what would be expected based on the national average.

We can never know with absolute certainty which programs are truly underperforming and which observations might be due to random chance. The Bayesian method calculates a ''best guess'' along with a plausible range for each program's hazard ratio (HR), allowing for probability statements regarding where the program's HR is likely to be (the HR is analogous to the O/E ratio used in the current system). For example, we could say, ''We are 75% certain that this program's death rates are more than 20% higher than expected (i.e. $HR > 1.20$).'' The MPSC may decide to review a program if that program's estimated HR exceeds a certain clinical threshold; for example, 1.20. However, this may result in identifying a program for review with only 50% certainty that its HR exceeds 1.20. Alternatively, it is possible to incorporate the probability that a program's HR exceeds the threshold into the decision-making process. For example, the MPSC could use a probability threshold of 75% and decide to review a program if certainty is more than 75% that its HR is above 1.20.

In the Bayesian context, it is difficult for small-volume programs to provide enough evidence to achieve the 75% probability threshold. The 75% threshold indicates strong evidence for poor performance, and small programs do not provide enough data to indicate strong evidence for poor performance or for good performance. Therefore, one could also incorporate an additional clinical threshold to indicate a nonnegligible chance of very poor performance. For example, one could flag a program if probability is more than 10% that its mortality rate is more than 150% worse than expected, that is, an HR greater than 2.5. An example flagging system, as shown in Figure 1, would flag a program for further review if the probability exceeds 75% that the program's HR is greater than 20% higher than expected (left panel) or if the probability exceeds 10% that the program's HR is greater than 150% higher than expected (right panel). Following discussions with the members of the MPSC in 2012 and 2013, the MPSC supported pursuing two complementary flagging criteria to identify centers with either (1) strong evidence of underperformance or (2) nonnegligible evidence of strong underperformance.

The goals of this study were:

1. To determine optimal flagging thresholds using the Bayesian methodology to achieve the stated goals of maximizing true positives while holding false positive flagging rates to approximately 5% regardless of program volume;
2. To compare the optimal Bayesian flagging approach to the results of the current flagging system on the most recently available PSR results from July 2012; and
3. To assess the optimal Bayesian flagging approach using a historical group of flagged programs from the January 2007, July 2007 and January 2008 PSR cycles with known actions initiated by the MPSC as a result of flagging under current SRTR methods. This allows assessment of true and false positive rates of flagging using Bayesian methods compared with the traditional method using previously flagged programs with resulting MPSC actions known.

Methods

Study populations

We used data from OPTN and SRTR. The SRTR data system includes data on all donors, waitlisted candidates and transplant recipients in the United States, submitted by the members of OPTN, and has been described elsewhere (6). The Health Resources and Services Administration, US Department of Health and Human Services, provides oversight of the activities of the OPTN and SRTR contractors.

Simulation study

The goal of the simulation study was to find optimal flagging boundaries for each of the two flagging criteria (strong evidence of underperformance, or nonnegligible evidence of strong underperformance). To meet this goal, we simulated 57 915 potential flagging algorithms using the Bayesian methodology. Four flagging thresholds were modified in each of the simulations:

1. The HR ratio threshold (HR_1) to be used for the ''strong evidence of underperformance'' criterion (e.g. strong evidence that the HR is greater than HR_1; an example is 1.20 as shown in Figure 1).
2. The probability threshold (P_1) to be used for the ''strong evidence'' criterion (e.g. the probability that the program's true HR is greater than HR_1 is greater than P_1; an example is 75% as shown in Figure 1).
3. The HR ratio threshold (HR_2) to be used for the ''nonnegligible evidence of strong underperformance'' criterion (e.g. nonnegligible evidence that the HR is greater than HR_2; an example is 2.50 as shown in Figure 1).
4. The probability threshold (P_2) to be used for the ''nonnegligible evidence of strong underperformance'' criterion (e.g. the probability that the program's true HR is greater than HR_2 is greater than P_2; an example is 10% as shown in Figure 1).

We first simulated patient deaths for all kidney, liver, heart and lung transplant programs with expected 1-year adult patient deaths estimated in the July 2012 PSR, assuming that each program was performing as expected, 2500 times. We then simulated patient deaths for the same programs, assuming that their patient death rates were two times their expected rates, 2500 times. Within each of these 5000 simulated data sets, we applied each of the 57 915 possible Bayesian flagging thresholds as

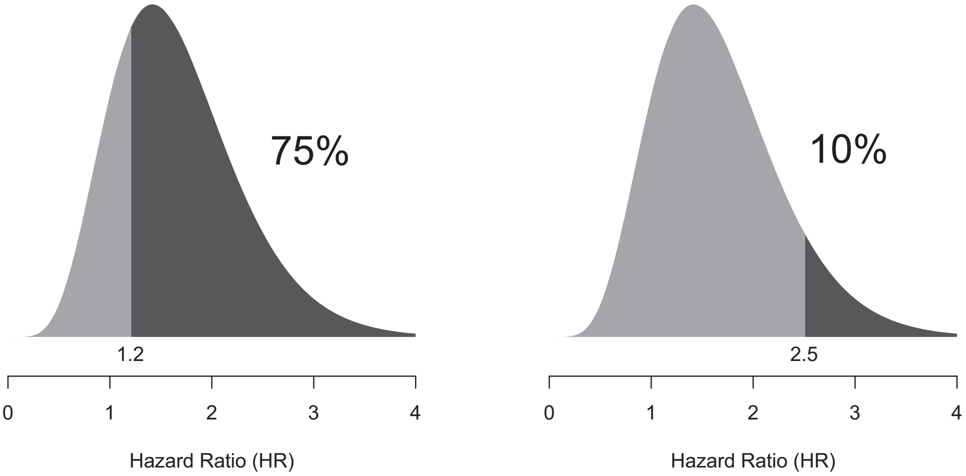

Figure 1: An optimal Bayesian flagging algorithm for underperformance based on 1-year patient death. According to this algorithm, a program would be flagged if probability was 75% that the program's patient death rate was 20% higher than expected (left panel), or if probability was 10% that the program's patient death rate was 2.5 times higher than expected (right panel).

described in items 1–4 above and calculated the true positive and false positive flagging rates.

We then scored each flagging algorithm by assessing a 0.05-point penalty for every one percentage point the false positive rate was above or below the 5% level, and a 0.01-point penalty for every one percentage point the true positive rate was less than 100%. This scoring algorithm prioritizes a 5% false positive rate over a 100% true positive rate by placing five times the penalty on a false positive rate that deviates from 5%. After we calculated the score for each of the 57 915 scoring algorithms, we ranked the algorithms, and the best algorithm was the one that optimized the false positive and true positive rates, keeping the false positive rate at approximately 5% across the range of program volumes while maximizing the true positive rate.

Comparison of the optimal flagging algorithm to July 2012 PSR results

Once the simulation determined the optimal flagging algorithm, we compared rates of flagging using current versus Bayesian methods for kidney, liver, heart and lung transplants occurring January 1, 2010, through December 31, 2011, and released as SRTR PSR in July 2012. This allowed us to compare the types of programs that would have been flagged had the new Bayesian system been in place for the July 2012 PSR cycle.

Comparison of the optimal flagging algorithm with previous MPSC actions

To assess how the Bayesian algorithm would have performed at identifying programs that were deemed to be truly in need of intervention, we used actions initiated by the MPSC based on current flagging criteria for transplant programs in the 12 months after three program-specific reporting cohorts:

1. July 1, 2003, through June 31, 2006, data released January 2007.
2. January 1, 2004, through December 31, 2006, data released July 2007.
3. July 1, 2004, through June 31, 2007, data released January 2008.

For each of these three cohorts, United Network for Organ Sharing (UNOS) staff classified flagged programs into ''true positives'' and ''false positives'' based on whether a significant transplant program event took place following MPSC review. This is not necessarily a ''gold standard'' analysis such as we have in the simulation study, but it does allow us to assess how the Bayesian algorithm would perform at flagging or not flagging programs in these cohorts with recorded MPSC actions. Significant transplant program events were defined as the program withdrawing from OPTN, prompting peer visits, becoming an OPTN member not in good standing or prompting informal discussions or other inquiries by the Performance Analysis and Improvement Subcommittee (PAIS) of the MPSC. If no significant events occurred and the program was no longer under review by the PAIS 1 year after flagging, the flagging was considered to be a false positive, and vice versa.

Results

Identifying an optimal flagging boundary through a simulation study

After we simulated 57 915 different algorithms, the current flagging algorithm ranked 11 987th based on the scoring system minimizing false positive and maximizing true positive flags (Figure 2). The best performing Bayesian algorithm (Figure 1) flagged a program if either of the following two criteria were met:

1. Strong evidence of underperformance: The probability that the HR was above 1.20 was greater than 75%, or
2. Nonnegligible evidence of strong underperformance: The probability that the HR was above 2.50 was greater than 10% (Table 1).

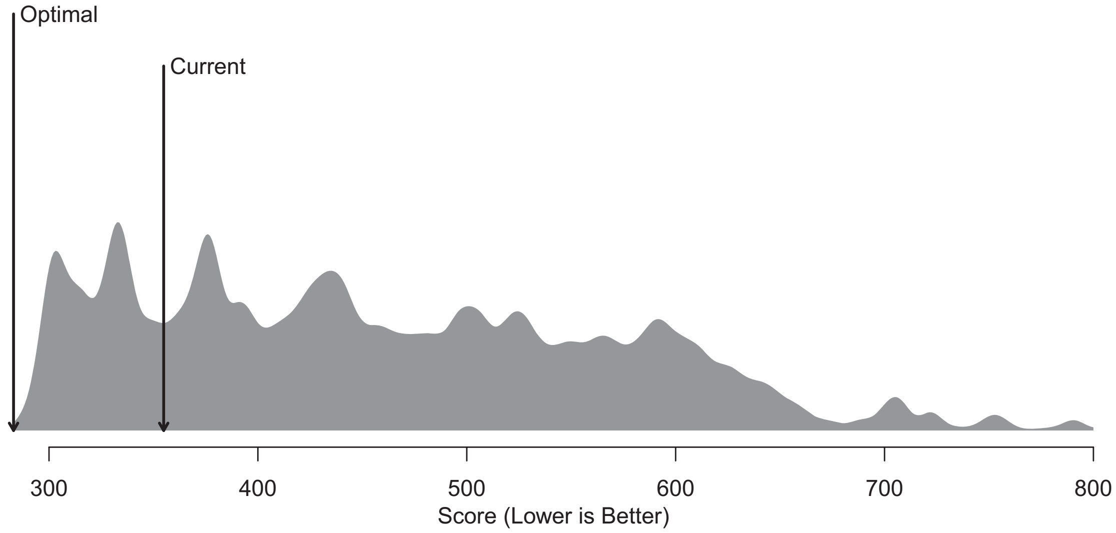

Figure 2: Distribution of algorithm scores. Lower scores indicate lower false positive and higher true positive flagging compared with actual flagging in the July 2012 program-specific report cohort. Of the algorithm scores, 93.7% were less than 800.

Table 1: Ranking of the top five algorithms and the current flagging algorithm out of 57 915 simulated flagging algorithms based on scores[1]

Rank	Hazard ratio 1	Probability 1	Hazard ratio 2	Probability 2	Score[2]
1	1.20	0.75	2.50	0.10	283.0
2	1.20	0.75	2.25	0.15	283.1
3	1.20	0.75	2.90	0.05	283.3
4	1.25	0.70	2.50	0.10	283.4
5	1.20	0.75	2.45	0.10	283.6
11 987		Current flagging algorithm			354.9

[1] Algorithms were based on meeting Bayesian probabilities of exceeding either of two threshold hazards for patient death (see text for details).

[2] Scores were based on points to minimize false positives and maximize true positives (see text for details).

The best performing Bayesian flagging algorithm achieved the stated goal of a false positive rate of approximately 5% across the range of program volumes (Figure 3, upper panel), while maximizing the ability to identify true positives (lower panel). This compared favorably with the current flagging algorithm (Figure 4). Notably, fewer small-volume and more mid-volume program true positives were identified with the best performing Bayesian flagging algorithm compared with current methods (Figure 5).

A comparison of the optimal Bayesian algorithm with current flagging methods

Compared with the number of programs flagged using the current algorithm in the July 2012 program-specific reporting period, the total number of programs flagged was 102 under the current system and 97 with the best performing Bayesian algorithm (Table 2). However, the number of small-volume programs flagged with the Bayesian algorithm declined while the number of programs

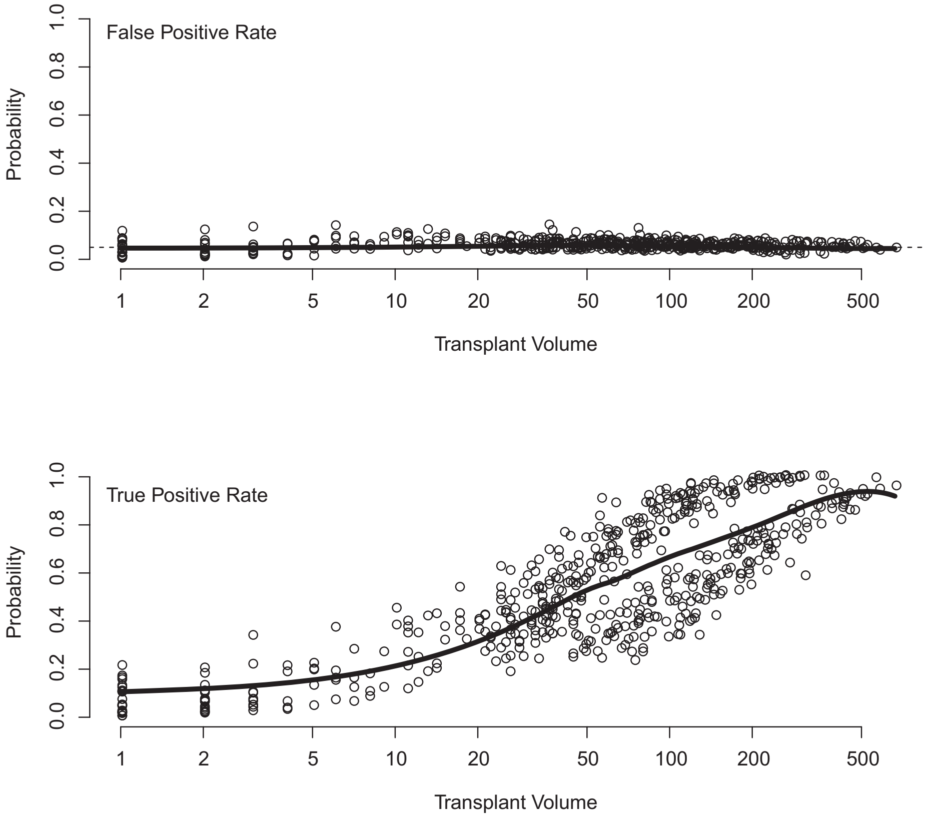

Figure 3: Bayesian flagging method; false positives (upper panel) and true positives (lower panel). The best performing Bayesian method was used: programs whose probability of a hazard ratio above 1.20 is greater than 75%, or whose probability of a hazard ratio above 2.50 is greater than 10%.

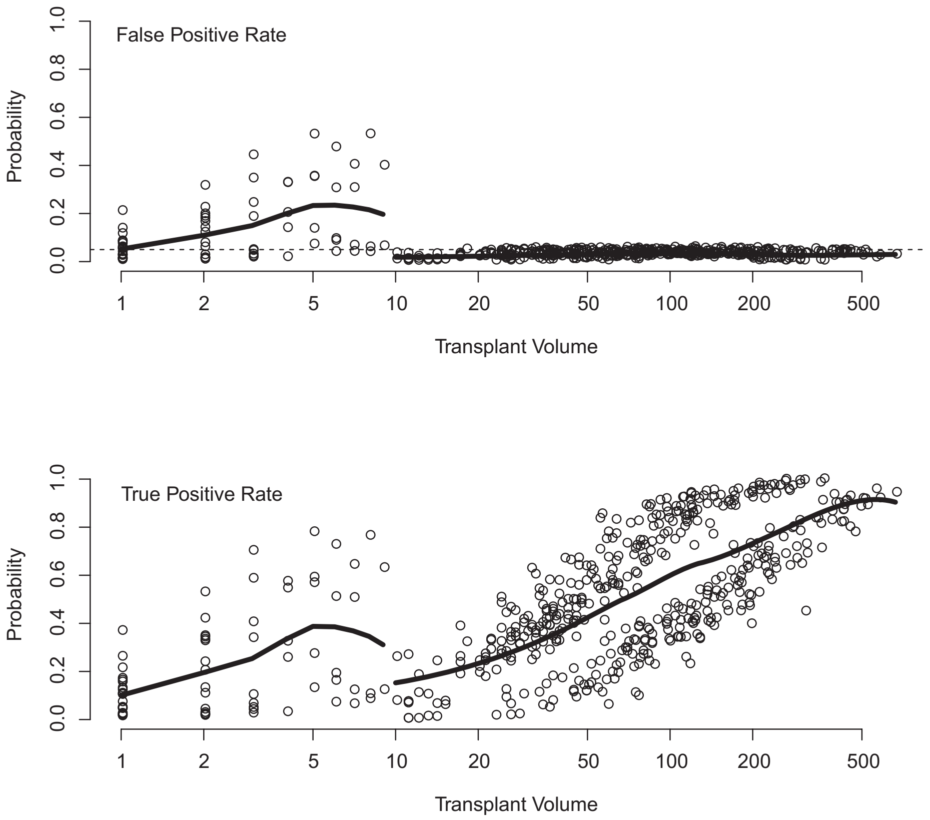

Figure 4: Current flagging method; false positives (upper panel) and true positives (lower panel).

in the mid-volume range increased, as predicted by the simulation study. Simulations also suggest that mid-to-large-volume programs flagged using the Bayesian system are more likely to be true positives than under the current flagging system.

A comparison of the optimal flagging algorithm with previous MPSC actions

Using historical data on MPSC final actions for flagged programs in 2007–2008, we compared whether or not actions were taken by the MPSC in response to programs

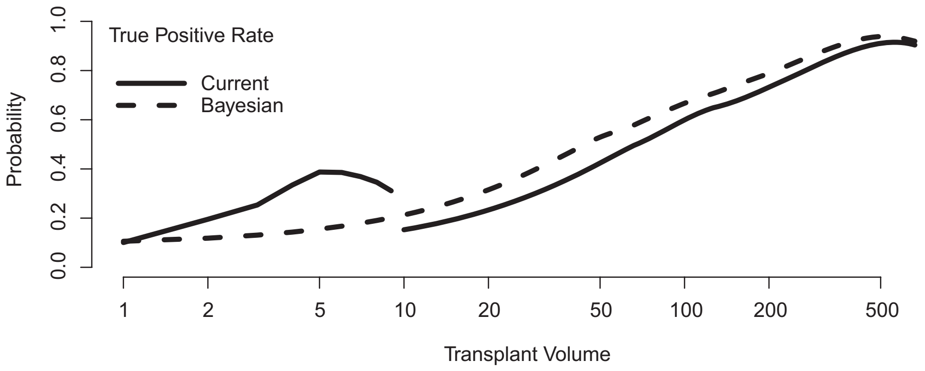

Figure 5: Comparison of the best performing Bayesian method versus the current method of flagging programs of different volumes. The best performing Bayesian method was used: programs whose probability of a hazard ratio above 1.20 is greater than 75%, or whose probability of a hazard ratio above 2.50 is greater than 10%.

Table 2: Number of transplant programs flagged by an optimal Bayesian algorithm versus the current flagging algorithm[1]

			Number flagged	
Program volume[2]	Number of programs[3]	Total number of transplants[4]	Current[5]	Optimal Bayesian[6]
1–9	223	799	54	15
10–49	270	7519	22	44
50–99	126	9139	11	19
100–249	147	23 694	11	15
250–744	61	23 977	4	4
Any (1–744)	827	65 128	102	97

[1] Based on data from program-specific reports released July 2012.
[2] Number of transplants January 1, 2010, through December 31, 2011.
[3] Number of kidney, liver, heart and lung programs.
[4] Total number of transplants in all programs of that volume.
[5] Number flagged by the current methods.
[6] Number flagged by an optimal Bayesian method (see text for details).

Table 3: Historical analysis of Bayesian algorithm performance for programs reviewed by the Membership and Professional Standards Committee (MPSC), January 2007–January 2008

Program	No action[1] (false positive)	Bayesian algorithm would have flagged	Action taken[1] (true positive)	Bayesian algorithm would have flagged
Kidney	82	41/82 (50%)	73	70/73 (96%)
Liver	51	18/51 (35%)	38	27/38 (71%)
Heart	56	24/56 (43%)	43	25/43 (58%)
Lung	21	12/21 (57%)	18	15/18 (83%)

[1] Actions taken by the MPSC included: requiring program withdrawal from the Organ Procurement and Transplantation Network, requiring peer visits, defining the program as a member not in good standing, or requiring informal discussions or other inquiries.

being flagged by current methods with hypothetical flagging by the best performing Bayesian flagging algorithm (Table 3). Results suggested that the best performing Bayesian method would result in many fewer false positives at the expense of missing some true positives.

Discussion

Identifying transplant programs with lower than expected outcomes is difficult, given large differences in program volumes and the resulting heterogeneity in statistical power. Not surprisingly, the current method of comparing observed-to-expected outcomes flags a disproportionate number of small programs. Not only is this inherently unfair to the smaller programs, but it also forces the MPSC to expend more time and effort detecting potential problems that affect relatively few patients than detecting potential problems and flagging relatively more of the larger programs.

The results of the current analysis demonstrate that a Bayesian method can be used to shift the preponderance of programs flagged from small- to mid-volume programs (Table 2). Simulations suggest that the majority of flags in the small-volume range are false positives, and the historical analysis of programs flagged by the MPSC supports this finding (Table 3). Furthermore, simulations suggested that the current flagging algorithm fails to identify true positives in the mid-volume range (Figure 5). The Bayesian analyses divert attention from small-volume programs, avoiding many of the false positives, and shift attention to mid-volume programs to detect true positives that affect a larger number of patients.

Since underperforming transplant programs are relatively few, and flagging false positives is costly, a low false positive rate is arguably desirable. However, because we can never achieve 0% false positives and 100% true positives, we identified the optimal Bayesian flagging boundary that resulted in a 5% false positive rate across the range of program volumes while maximizing the true positives.

The Bayesian method replaces ''all or none'' thresholds with *probabilities* for exceeding thresholds that define underperformance. By setting two probability thresholds for underperformance, one to identify strong evidence of underperformance (HR greater than 1.2 with greater than 75% probability), and another to identify nonnegligible evidence of very bad performance (HR greater than 2.5 with greater than 10% probability, Figure 1), fewer small programs will be flagged, and false positives resulting from low volume and low statistical power will be avoided.

However, even an optimal Bayesian method cannot completely overcome the problems inherent in detecting poorly performing small-volume programs. Therefore, regulatory bodies such as MPSC and CMS may opt to randomly audit small-volume programs or to use the current methods or other methods for flagging them.

In summary, SRTR has adapted Bayesian methods for analysis of SRTR PSR and will transition public reporting of transplant outcomes to use the Bayesian methods. Bayesian methods have the major advantage of flagging programs of different volumes more evenly than the methods currently being used. Bayesian methods also add an appealing quantitative approach to identifying underperforming programs by calculating the probability that programs exceed a predetermined threshold, rather than using an all-or-none p-value approach.

Acknowledgments

These data were presented in part at the American Transplant Congress, May 22, 2013, and published in abstract form (7). This work was conducted under the auspices of the Minneapolis Medical Research Foundation, contractor for SRTR, as a deliverable under contract no. HHSH250201000018C (US Department of Health and Human Services, Health Resources and Services Administration, Healthcare Systems Bureau, Division of Transplantation) and UNOS, contractor for OPTN, under contract no. 234-2005-370011C. As a US 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 US Government, UNOS or OPTN. The authors thank SRTR colleagues Delaney Berrini, BS, for manuscript preparation and Nan Booth, MSW, MPH, ELS, for manuscript editing. The authors also thank the members of the SRTR Technical Advisory Committee for their guidance, and the anonymous editors and reviewers of earlier drafts of the manuscript for their many helpful comments.

Disclosure

The authors of this manuscript have no conflicts of interest to disclose as described by the *American Journal of Transplantation*.

References

1. National Organ Transplantation Act of 1984. Pubic Law 98-507, Title 42-Section 273. 10-19-0984. 98 Stat. 2339-2348.
2. Dickinson DM, Arrington CJ, Fant G, et al. SRTR program-specific reports on outcomes: A guide for the new reader. Am J Transplant 2008; 8: 1012–1026.
3. Kasiske BL, McBride MA, Cornell DL, et al. Report of a consensus conference on transplant program quality and surveillance. Am J Transplant 2012; 12: 1988–1996.
4. Ash AS, Fienberg SE, Louis TA, Normand S-LT, Stukel TA, Utts J. Statistical issues in assessing hospital performance. 2012. Available at: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf. Accessed December 6, 2013.
5. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. Ann Intern Med 1997; 127: 764–768.
6. Leppke S, Leighton T, Zaun D, et al. Scientific Registry of Transplant Recipients: Collecting, analyzing, and reporting data on transplantation in the United States. Transplant Rev 2013; 27: 50–56.
7. Salkowski N, Edwards E, Leighton T, Israni A, Kasiske B, Snyder J. Improving the SRTR methodology used to identify potentially underperforming transplant programs in the United States [abstract]. Am J Transplant 2013; 13(S5): 199–200.