American Journal of Transplantation 2008; 8 (Part 2): 1012–1026
Blackwell Munksgaard

doi: 10.1111/j.1600-6143.2008.02178.x

SRTR Program-Specific Reports on Outcomes: A Guide for the New Reader

D. M. Dickinson[a,*], C. J. Arrington[a], G. Fant[b], G. N. Levine[a], D. E. Schaubel[c], T. L. Pruett[d], M. S. Roberts[e] and R. A. Wolfe[a]

[a] *Scientific Registry of Transplant Recipients, Arbor Research Collaborative for Health, Ann Arbor, MI*
[b] *Health Resources and Services Administration, Rockville, MD*
[c] *Scientific Registry of Transplant Recipients/University of Michigan, Ann Arbor, MI*
[d] *University of Virginia, Division of Transplantation, Charlottesville, VA*
[e] *University of Pittsburgh School of Medicine, Pittsburgh, PA*
* *Corresponding author: David M. Dickinson, david.dickinson@ArborResearch.org*

Differences in outcomes indeed exist among transplant programs and organ procurement organizations (OPO). A growing set of tools are available from the Scientific Registry of Transplant Recipients (SRTR) to measure and assess these outcomes in the different phases of the transplant process. These tools are not intended to compare two individual programs, rather to help identify programs whose practices may need further scrutiny, to be either avoided, corrected or emulated.

To understand which differences in outcomes might be due to underlying differences in populations served and which might be due to differences in treatment, it is important to compare outcomes to 'risk-adjusted' expected values. Further, it is important to recognize and assess the role that random chance may play in these outcomes by considering the p-value or confidence interval of each estimate. We present the reader with a basic explanation of these tools and their interpretation in the context of reading the SRTR Program-Specific Reports.

We describe the intended audience of these reports, including patients, monitoring and process improvement bodies, payers and others such as the media. Use of these statistics in a way that reflects a basic understanding of these concepts and their limitations is beneficial for all audiences.

Key words: Program-specific, risk-adjustment, SRTR, survival, transplant outcomes

Introduction

The Transplant Program- and OPO-Specific Reports published by the Scientific Registry of Transplant Recipients (SRTR) are intended to help evaluate whether the nation's organ transplant system is efficient at providing optimal care to patients, especially given the constraints on the precious resource of donor organs. These reports provide a range of performance outcomes at the nearly 1000 transplant programs and organ procurement organizations (OPOs) that participate in the Organ Procurement and Transplantation Network (OPTN). Their public availability provides the opportunity for any audience to see which transplant programs are not transplanting candidates as quickly as expected, which programs have high posttransplant survival rates and which OPOs recover different types of organs at the highest rates.

The reports themselves are part of a whole family of tools, consistent in their statistical approach, that are used by multiple audiences to help understand the outcomes achieved in different phases of the transplant process. As the statistics included in these reports are used increasingly by various audiences—the OPTN Membership and Professional Standards Committee, CMS and private insurers, the popular press and patients and families—many readers have requested a closer look at their interpretation. This article takes the new reader of these reports through many of the statistics included and addresses some of the most common questions raised about these reports.

Hereafter, we generally refer to these reports as 'Program-Specific Reports' unless specifying those for transplant programs or OPOs.

We aim to explain basic themes included in these reports to readers who are new to these statistical concepts, and provide an overview for readers who are familiar with the concepts but not with their application in this area. For readers seeking more technical detail, further explanations have been previously published (1) and are also available in the 'Background and Methods' section of the Program-Specific Reports at www.ustransplant.org.

We start by surveying the range of statistics that describe the different parts of the transplant process, then address the following questions:

(i) Are there really differences among transplant programs and OPOs? In types of performance measurements?
(ii) Are the differences we observe among programs merely due to differences in patients, donor organs or random chance? What role does 'risk adjustment' play in helping us distinguish these from differences that can be attributed to the transplant program or OPO?
(iii) Who are the audiences for the reports, and how do they use the reports for performance measurement?
(iv) How should these measures affect transplant program or OPO behavior?

Which Statistics and Parts of the Transplant Process Are Covered?

The organ procurement and transplantation process is complex with many trade-offs in outcomes. The Program-Specific Reports present characteristics and outcomes of multiple stages of the transplant process, without value judgment about these trade-offs. For example, a transplant center may be able to transplant a high number of its wait-listed patients, at the expense of poorer transplant outcomes that arise from the use of less ideal organs. However, a donor organ with many donor comorbidities may be better for a given candidate than waiting and perhaps receiving none at all.

Figure 1 gives an overview of these stages and the measures presented in the reports, intended to help evaluate the efficient operation of the transplant system and effective use of a limited supply of organs to answer a growing demand.

The outcomes listed in the top left box, 'Procurement of Donor Organs', help us answer questions pertaining to the first stage in the process defining efficient use and optimal patient care: identification and procurement of donor organs. An OPO and the hospitals with which it works are primarily responsible for identifying eligible donors from among the in-hospital deaths, working to ensure that appropriate candidates become donors, and seeing that, if possible, all of the available organs from a procured donor are used. Measures of these three outcomes are shown in the OPO-specific reports, and the implementation of currently planned expansion of data collection for eligible donors will help refine these measures.

The top right box, 'Care for Patients Awaiting Organs', also reflects whether a limited supply of available organs is being used efficiently, and how close that supply meets the demand, this time more from a patient's point of view. Though the transplant program may not be the primary direct caregiver before the transplant takes place, characteristics and practices at the center and within the donation service area (DSA) may influence how long the patient waits for a suitable organ, and the patient's chances for survival during that time. The primary waiting list outcomes—transplant rate and mortality rate—may indicate a center's willingness to accept higher risk organs, the OPO's ability to find organs or even the rate at which potential organ donors become available.

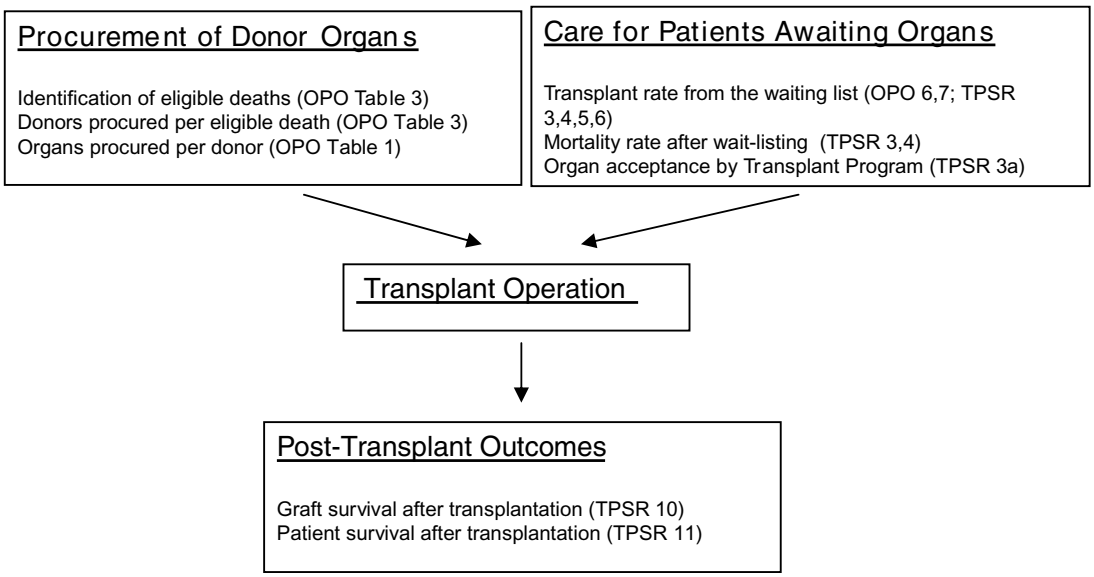


Figure 1: Phases and measures presented in the Program-Specific Reports.

The bottom box, 'Post-Transplant Outcomes', addresses questions regarding the last stage of the transplant process by measuring posttransplant graft and patient survival. Outcomes at 1 month, 1 year and 3 years after transplant all reflect the risk of the transplant operation, as well as immediate- and long-term follow-up care.

In addition to the outcomes listed in Figure 1, the reports also give descriptive background about the populations served by each program or OPO. For transplant programs, different tables show profiles of the wait-listed patients and transplant recipients, as well as profiles of the donors chosen for transplant. For OPOs, different tables describe the geographic areas served by the OPO, the characteristics of the donors and organs procured and which transplant programs eventually use organs recovered by the OPO. These descriptive tables provide background to the outcomes tables that we focus on here.

Why Program- or OPO-Specific? Are There Really Differences Among Programs?

Differences do indeed exist among programs. As an example, Figure 2 shows unadjusted 1-year kidney graft survival after deceased or living donor transplant. The heights of the vertical bars represent the number of transplant failures per 10 transplants performed at each of the 238 different kidney transplant programs in the country. Six programs, at the left end of the figure, had more than two failures per 10 transplants, corresponding to 1-year graft survival of less than 80%. At the other end, 24 programs had no failures among their transplants.

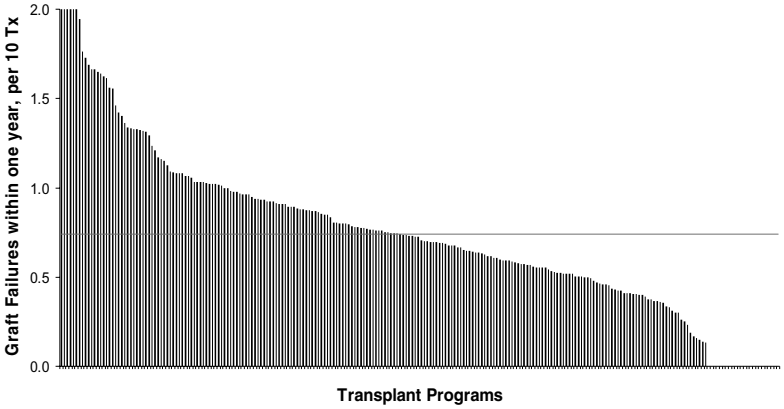
In between, we see wide variation of transplant survival around the horizontal line indicating the national average of 0.74 failures per 10 transplants, or about 93% survival. Fully one in four programs experience graft survival at or below 90%, while another quarter of programs experience 1-year graft survival rates above 95%. There is a fairly large difference between the top quarter of programs and the bottom quarter.

Similar variation exists among all of the outcomes listed in Figure 1, and for graft and patient survival for other organs.

One must be careful in interpreting such differences, because these differences could arise from several factors:

(i) Differences among the patients served by the center: for example, one center may treat patients who, on average, are older or more severely ill than most other centers, and therefore expect lower survival.
(ii) Differences in the characteristics of the donors available**:** some donor organs are higher quality than others, and some centers may have access to donors that are a better match for its patients—or simply to more donors overall.
(iii) Differences in treatment practices**:** some transplant teams may be more experienced or skillful, or otherwise provide better care both during and after the transplant operation. The care received outside of the transplant center may also be of variable quality.
(iv) Random variation**:** sometimes centers experience poor, or excellent, outcomes just as a matter of random chance, particularly for smaller programs.

To disentangle the differences of interest from those that might only confuse the question, we use a process called 'risk adjustment', which we describe in the next section. In the case of determining which programs provide excellent care or need improvement, the differences of interest are those due to variations in patient treatment; we will risk-adjust for measured and reported patient and donor characteristics that might otherwise confuse our results, and thereby exclude these differences from our comparisons.



Source: *SRTR Transplant Program Specific Reports, published July 2007. Includes deceased and living donor transplants performed in the 2.5 years ending 6/30/2006.*

Figure 2: Unadjusted graft failures per transplant.

What Does Risk Adjustment Mean?

Risk adjustment is a statistical method through which we can factor out differences in outcomes that reflect differences among patients or donors in order to infer the magnitude of differences due to treatment practices. The use of risk adjustments allows us to answer questions such as:

> 'What is the survival we observe at this program, and how does it compare to what we would expect based on typical results for the types of patients served and donors used'?

This question has two important features. First, the main focus is on the 'simple' actual survival rate achieved and observed for the transplant program. However, this framing of the question also emphasizes that the actual survival rate must be compared to what would be expected for the 'case mix' of patients served and donors used, in order to avoid attributing poor (or good) outcomes that are associated with this case mix to the program's treatment practices instead.

Risk adjustment is a quantitative or statistical process that helps us to account for differences in patients and donors in making a more valid comparison of other factors, such as treatment practices. The next several paragraphs explain the use of risk adjustment with practical examples from the transplant community.

A simplified example: adjusting for recipient age

To help understand how we calculate expected outcomes, and to illustrate why it is important to do so, consider the example of how the ages of the transplant recipients may affect a transplant program's outcomes. In this simplified illustration, we will assume that all patients and donors are alike, except that they have different ages. Figure 3 shows our calculation, beginning with the national observed patient survival for two adult age groups, taken from the 2006 OPTN/SRTR Annual Report. Adults under age 65 had much higher survival, 79% at 3 years, than did those over 65, 71% at 3 years following transplant. Because only one in 10 adult liver recipients was over 65, the national average survival rate was very close to the rate for these younger recipients, at 78% (which is the 'weighted average' between the 79% for 90% of recipients and 71% for the other 10%).

Unlike the nation as a whole, the sample transplant program shown in the three columns to the right treats mostly patients older than 65, or 90 of the 100 patients transplanted. Nine of 10, or 90%, of their younger patients survived, as did 66 of 90 (73%) of their older patients. Both groups of patients at this center did better than the corresponding groups in the nation, yet a simple comparison of the overall survival rate shows this center, at 75%, falling below the national average of 78%.

The lower panel shows the calculation of the expected survival rate, which allows us to make a proper comparison between program-specific and national experience. For the 10 younger patients, we can expect a 79% survival rate based on the national experience, or that 7.9 patients should survive; for the 90 older patients, we expect 71% or 63.9 to survive. Together, we expect 71.8 survivors or 71.8% of the total 100 transplants. The favorable comparison of the center's survival of 75% to the expected survival of 71.8% is consistent with the results we find for each age group. Had we not used risk adjustment for age to account for the older patients served by this center, we might have improperly characterized this program as under-performing with respect to patient survival, when in fact they are actually performing better than expected in treating a more challenging patient mix.

Risk adjustment by within-group comparison

OPO-specific report Table 1 measures both the number of organs recovered for transplant and the number transplanted, both expressed as 'per donor'. This table illustrates simple risk adjustment by comparison within group, as shown in the example of age-adjustment above. The average number of organs per donor is calculated separately for standard criteria donors (SCD) as well as expanded criteria donors (ECD) and donors after cardiac death (DCD). By reporting these averages separately for each donor type, we can compare the organ yield for an OPO against a national average for similar donors. Just as in the age example above, to fairly characterize the per donor count of organs for an OPO that procures organs from many ECDs, donor-type specific comparisons must be made because

	National		Transplant Program				
	% in group	% survival		Transplants		Survivors	Survival %
Observed							
Age 18-64	90%	79%		10		9	90.0%
Age 65+	10%	71%		90		66	73.3%
Overall		78%					75.0%
Expected							
Age 18-64		79%	X	10	=	7.9	
Age 65+		71%	X	90	=	63.7	
Overall						71.6	71.6%

Figure 3: Simple age adjustment for 3-year liver survival.

Source: *Analysis based on 2006 OPTN/SRTR Annual Report, Table 9.12a*

Table 1: Effect of expanded criteria donor definition components on kidney graft survival

Factor	Hazard ratio
Hypertension	1.18
Creatinine (per 1 mg/dL)	1.10
Donor age: 65+ (ref = 35–49)	1.49
COD stroke (vs. head trauma)	1.31
'ECD' classification	1.08

Calculated as exp (Beta) from 1-year kidney graft survival model, CSRs released 01/11/2007.
Source: SRTR.

the lower yield among ECDs may detrimentally decrease the overall average for the OPO that procures more ECDs than average.

Similarly, report Table 5 of the transplant Program-Specific Reports adjusts measurements of waiting time until transplant by analyzing this calculation separately for wait-listed patients with different characteristics. The characteristics chosen include those that may likely influence the time until transplant, such as disease severity or immune sensitivity.

Risk adjustment for multiple factors by regression model

In the Program-Specific Reports, the tables showing graft and patient survival, waiting list outcomes, and donation rates use a more complex adjustment method to account for the many patient and donor factors that should be included in the risk adjustment. To simultaneously adjust for a long list of factors in the same way that age is controlled for above, the SRTR uses a statistical technique called the Cox regression model (2).

The Cox model uses observations of all the patients and donors in the country, and their characteristics and outcomes, to estimate the effect of each characteristic on outcomes. We then apply these estimated effects to each patient-donor combination, allowing us to calculate an expected outcome for each patient, which can be added together for all patients treated by a transplant program. This effect is how each factor is 'weighted' in the risk-adjustment process. A broader description of the factors selected follows later in this article.

For example, many programs use ECD kidneys for recipients who are likely to die before having the opportunity to receive a non-ECD kidney. To ensure that the lower survival rate associated with these donors does not, on its own, indicate poor performance for the transplant program as a whole, we incorporate these donor factors into the models for expected survival. Table 1 shows the factors used in identifying an ECD kidney and their separate effects on 1-year graft survival. Not all ECD kidneys are characterized by all of these factors. A kidney from a donor with a history of hypertension, whether classified as ECD or not, carries with it a risk of graft failure 1.18 times that of an organ from a donor without hypertension, or 18% higher risk (Table 1). That is, if a patient with an organ from a nonhypertensive donor had a probability of graft loss at 1 year of 0.05, then the same patient with a hypertensive—but otherwise similar—donor would have a probability of graft loss at 1 year of $0.05 \times 1.18 = 0.059$. If that same donor were also older than 65, the kidney would be another 1.49 times as likely to fail compared to a donor between 35 and 49, for total elevated risk of $1.18 \times 1.49 = 1.76$. By multiplying the hazard ratios listed, note that a kidney from a donor with all of the characteristics listed in Table 1 represents a graft failure risk more than three times that of a kidney from a donor with none of these characteristics.

Reading and interpreting a risk-adjusted table

Figure 4 shows excerpts from a recent Program-Specific Report for a liver transplant center. As was the case in the simplified age example presented above, the reader should be careful to compare the center's outcomes to those expected based on risk adjustment, rather than the unadjusted national average.

In this particular example, the center achieves a posttransplant survival percentage at 1 year of 87.78% (line 3). This is higher than the national average of 86.26% (line 3), but lower than the expected survival of 89.41% (line 4). The following conclusions seem logical to make:

(i) The expected survival rate at this center is higher than the national average, suggesting that the types of patients treated at this facility, or the types of donors used by this facility, typically have above-average outcomes.
(ii) The survival rate we observe as achieved by this center, while higher than the national average, is not as high as might be expected given the typical results for similar patients in the rest of the nation.
(iii) Had no risk adjustment been performed, a comparison to the national average would be misleading.

This panel of the Program-Specific Reports table formulates statistics from the perspective of a user asking to estimate the chances that a patient would be living 1 year after transplant surgery, if transplanted at a specific center.

Standardized ratios

Lines 5–10 of the table in Figure 4 show another perspective that can be interpreted to find the fraction of excess events (or percentage shortage of events). The 'standardized ratio' is the number of observed events (deaths after transplant, graft failures, transplants, deaths after waitlisting or organ donors procured) divided by the number of those events that would be expected, according to the risk-adjustment model and time followed.

Line		Center 1 Year	National 1 Year
	Adult (Age 18+)		
1	Transplants (n=number)	90	10,781
2	**Percent (%) of Patients Surviving at End of Period**		
3	Observed at this Center	87.78	86.26
4	Expected, based on national experience	89.41	
5	**Deaths During Follow-up Period**		
6	Observed at this center	11	1,392
7	Expected, based on national experience	8.48	1,392
8	Ratio: Observed to Expected (O/E)	1.30	1.00
9	(95% Confidence Interval)	(0.65-2.32)	
10	P-value (2-sided), observed v. expected	0.469	
11	How does this center's survival compare to what is expected for similar patients?	Not Significantly Different (a)	
12	Percent retransplanted	5.5	4.4
13	Follow-up days reported by center (%)	91.7	93.9
14	Maximum Days of Follow-up (n)	365	365

Figure 4: Program-Specific Report table 11—patient survival after transplantation, Sample Liver Center.

Source: SRTR Program-Specific Reports, www.ustransplant.org

A standardized ratio equal to one indicates that the facility performed exactly as expected, given its case mix. Note that 'events' may be good (transplants from the waiting list, organ donors procured) or bad (deaths or graft failures). In the case of adverse events, such as the 'deaths during the follow-up period' shown in Figure 4, standardized ratios *above* one indicate worse than expected performance, while standardized ratios *below* one indicate better than expected performance. Conversely, for a good outcome such as transplant from the waiting list, the interpretation goes in the other direction: a standardized ratio above one would indicate better than expected performance.

Additionally, this ratio easily conveys the extent of the difference from the expected outcome, in this case the ratio of excess deaths. In Figure 4, 11 deaths were observed (line 6), compared to 8.48 expected (line 7), for a standardized ratio of 1.30. We observed 1.30 times as many deaths as we expected or 30% more. If the ratio were below one, it can be interpreted as the fraction of expected deaths; for example, a standardized death ratio of 0.80 indicates that we observe 0.80 deaths per expected death or 20% fewer (1–0.8 = 0.2).

Unlike the comparison of the percentage of patients surviving, the counts of observed and expected events used in the standardized ratio also take into account the timing of an event, and reflect an advantage for longer survival within the time period examined. For example, compare a patient who dies 1 day after transplant with a patient who dies 365 days after transplant. Each of these has the same effect on percentage surviving after 1 year, and each contributes one death to the numerator of the standardized ratio. Since the denominator of the ratio, the expected number of deaths, is much higher after 365 days, the adverse effect of the early death is weighted more heavily (3).

Limitations of risk adjustment

The risk-adjustment process can only account for differences among patients and donors that are measured and reported completely and accurately for patients across all transplant programs. Some characteristics that would likely impact posttransplant outcomes are not collected by the OPTN data collection process or the other sources used (3), because of the need to balance full adjustment with the burden of data collection. Though many clinical features affect patient survival, calculating—and adjusting for—the effect that those features 'usually' have on patients requires evidence from all programs and patients. Granular or even major differences in severity of illness or donor quality cannot be adjusted for if they are not reported uniformly across all centers. Examples might include the presence of interstitial fibrosis on the biopsy of a donated kidney or a measure of the severity of a recipient's coronary artery disease: though these factors surely affect survival, they are not reported in the data sources available.

Even when data are available, practices or characteristics that are exhibited only at a small number of programs are difficult to incorporate into the risk-adjustment process because no 'typical national experience' can be defined. It is therefore hard, or even impossible, to adjust for the risk of some new or experimental procedures because no 'typical' experience can be established from other programs.

Given the limitations of risk adjustment, as well as the role of random chance discussed later, it is useful to look at how

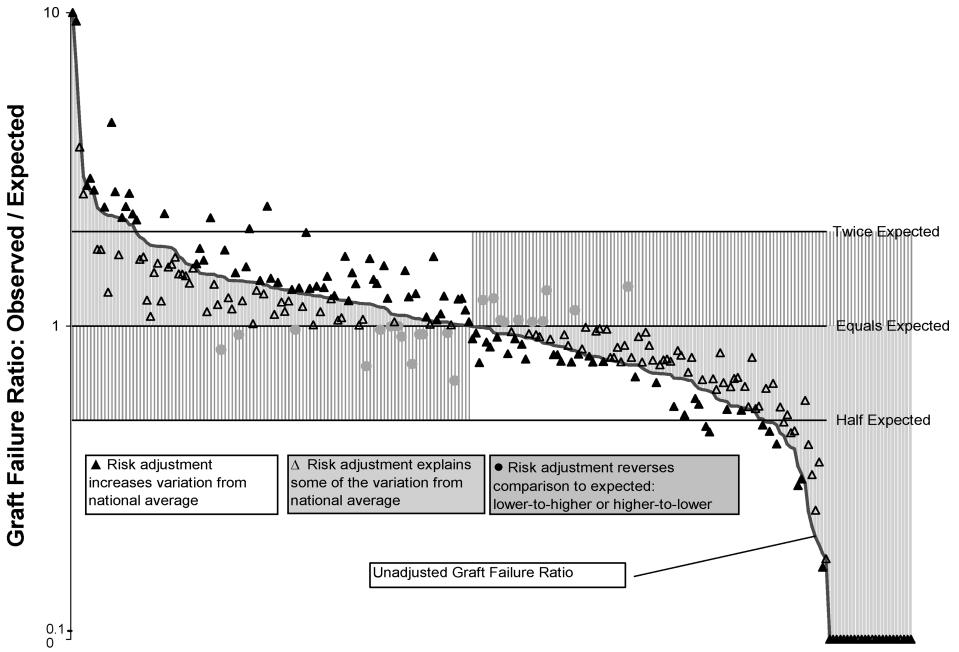
well the models we have perform. For survival models, the index of concordance (IOC) is used to measure how well the model predicts the outcomes. Specifically, of the number of patient pairs for which the ordering of the failure times is observed, the IOC is the percentage of observed orderings that are consistent with the orderings predicted by the model. The IOC can be interpreted as reflecting the percentage of variation in outcomes that is predicted by the factors included for risk adjustment; the remainder of the variation being due to the effects of treatment practices at the transplant program, as well as to unmeasured or unreported risk factors and random chance. For models released in January 2008, indexes of concordance range from 56% to 93%. The vast majority of the 44 models (for different organs, donor types and follow-up periods) had IOCs between 60% and 75%, leaving 25% to 40% of the variation due to unmeasured or unreported factors, treatment effects and random chance. IOCs are published along with the estimated effects of each risk adjustment factor in the PSR website section titled 'Risk Adjustment Models'.

After Risk Adjustment, Are There Still Differences Between Programs?

At the start of this article, we asked the question of whether there really are differences in outcomes among transplant programs. As exemplified by Figure 2, there can be tremendous variation in unadjusted outcomes. We now look at whether those differences 'disappear' when we look at adjusted statistics; whether the outcome differences we saw in Figure 2 are all due to patient and donor characteristics that can be risk-adjusted or 'controlled for in the model'.

In Figure 5, we show an unadjusted graft failure ratio, similar to the standardized ratio described above, but without case mix adjustment (the expected value is the simple national average, not depending on the types of patients served). This unadjusted ratio is the line in the background from the highest ratio (most graft failures per transplant) on the left to the fewest on the right, and mirrors the shape of the distribution shown in Figure 2. The horizontal line at observed/expected (O/E) = 1 identifies where each facility would be if it experienced exactly as many graft failures as expected. Note that the horizontal line represents, for the unadjusted survival, the same concept that it did as the national average in Figure 2: for an unadjusted statistic, the expected value is the national average.

Each dot (triangular or round) represents the standardized graft failure ratio for each program, risk-adjusted for case mix. For any center, the impact of risk adjustment is measured by the distance along a vertical line drawn through that center's adjusted value (dot), and its unadjusted value (line). After adjustment, about 45% of transplant programs move down (or up) into the shaded area of the graphic and



Source: SRTR Transplant Program Specific Reports, published July 2007. Includes transplants performed in the 2.5 years ending 6/30/2006.

Figure 5: Comparison of unadjusted to adjusted graft failure ratios among kidney transplant programs.

Table 2: Risk-adjusted tables, methods and factors in the SRTR Program-Specific Reports

Outcome concept	Report, table, name of statistic	Adjustment type	Adjustment factors (may vary with organ)
Identification of eligible deaths from among in-hospital deaths	OPO Table 3: 'Notification Rate'	Regression model with comparison to expected	Age, race, gender and cause of death among notifiable deaths
Donor conversion from eligible deaths	OPO Table 3: 'Donation Rate'	Regression model with comparison to expected	Notification rate, hospital characteristics such as presence of a trauma center. National data are not currently available on the characteristics of the eligible deaths.
Organs procured per donor	OPO Table 1: 'Organs Recovered/ Transplanted per Donor'	Within-group comparison	Comparison to national average separate for Standard Criteria Donor, Expanded Criteria Donors and Donation after Cardiac Death
Transplant rate from the waiting list	Transplant program report, Table 3: 'Transplant Rate'	Regression model with comparison to expected	Age, blood type, previous transplant, immunological sensitivity, primary disease, medical urgency status or disease severity
	Transplant program report, Table 4 and OPO Table 6: '% transplanted by a given time'	Within-group comparison	Age, blood type, previous transplant, immunological sensitivity, primary disease, medical urgency status or disease severity
Mortality rate after wait-listing	Transplant program report, Table 3: 'Mortality rate after wait-listing'	Regression model with comparison to expected	Age, blood type, previous transplant, immunological sensitivity, medical urgency status or disease severity
Posttransplant graft and patient survival	Transplant program report, Table 10: 'Graft Survival' 'Patient Survival'	Regression model with comparison to expected; within-group comparison (living vs. deceased donor; pediatric vs. adult)	Donor factors (demographics, history of related illnesses, cause of death, organ function measures); Recipient factors (demographics, disease severity, immunological sensitivity, other health status indicators, insurance); Donor-recipient match characteristics (antigen mismatches and blood type compatibility, cold ischemic time, compatibility of body size)

Source: SRTR.

toward the horizontal 'expected' line; these are transplant programs for which case mix factor explains a portion of the variation in performance.

For another 45% of transplant programs, the risk-adjusted ratio is even farther from one than its unadjusted counterpart; these are the darkened triangles. That is, differences in outcome among facilities may persist or even become more pronounced upon taking into account these patient and donor factors.

Sometimes the interpretation completely changes as a result of risk adjustment. These are the 10% of programs, indicated by round dots, which move all the way across the shaded area. For these, risk adjustment changes the overall interpretation from better performance to worse (or *vice versa*). Just as in the age example presented earlier, an accurate picture is obtained only by looking at risk-adjusted values.

Which Statistics Are Risk Adjusted?

Above, we outlined two different approaches to risk adjustment, both used in the SRTR OPO- and Transplant Program-Specific Reports: adjustment via a 'comparison within-group' or stratified analysis, and the use of a regression equation to adjust for multiple factors simultaneously. Table 2 lists which statistics in these reports are risk-adjusted, which method is used and overviews of the factors used for adjustment.

Note that adjustment factors differ for each organ. Moreover, the set of adjustment factors for a given organ may differ across reports, since the models are under frequent review by the SRTR. For a detailed list of current adjustment factors, see the risk-adjustment model description tables in the 'Background and Methodology' section of the Program-Specific Reports at www.ustransplant.org.

How Are the Factors for Risk Adjustment Determined?

Risk adjustment helps determine what results we would expect for 'similar' patients, according to the national experience. But what variables define similarity for our risk profile? The following list identifies risk factors that may impact posttransplant outcomes, and how likely they are to be appropriate for risk adjustment.

Patient characteristics? Almost always. Adjusting for patient characteristics helps ensure that centers are not penalized for treating patients who are more likely to have poor outcomes. For example, the age of the recipient is closely associated with outcomes, and not controlling for age might penalize centers that treat older patients, either as an explicit protocol element or due to the age distribution of the geographic area served.

Donor characteristics? Usually. Given a very short supply of organs available for transplant, programs and patients often decide to use an organ that is not optimal, determining that having this organ is better for the patient than remaining on the waiting list. Adjusting for a range of donor organ characteristics helps ensure that programs are not penalized for these decisions. For example, more and more centers are using ECD kidneys for patients who may be better off with those organs than lingering on the waiting list; by not controlling for these characteristics, which by definition result in elevated risk of graft failure, we would unfairly compare outcomes of ECD and non-ECD recipients, which might discourage the use of ECD organs.

Characteristics of the donor-recipient match? Much of the time. These characteristics include things like the blood type or antigen compatibility between the donor and recipient. Like other donor characteristics, transplant programs may need to accept organs that are imperfect matches. There is a trade-off between adjusting for compromises in choices about the donor, and adjusting for poor choices on the part of the transplant program.

Transplant center characteristics? Usually not. Center characteristics and practices may be associated with the differences that we are trying to identify in the Program-Specific Reports, and therefore should not be 'risk-adjusted away'. Program volume is a good example: even though larger programs may be associated with better outcomes, we want to give due credit to larger centers that perform well rather than adjusting away the differences associated with volume.

The SRTR updates these Program-Specific Reports every 6 months and, in the update process, incorporates ongoing enhancements to the risk-adjustment models. At each report, the risk adjustment is recalculated, and each year the SRTR focuses on reviewing the entire set of risk-adjustment covariates for one or more organs. Selection of model covariates is based on the entire body of analytical work performed by the SRTR for the OPTN committees and other groups, as well as input from OPTN committees. The following factors are considered:

What are the known predictors of survival? We focus on factors shown to be important either in SRTR analyses or in the medical literature. Factors are included when our analyses show that they have a low probability, less than 10%, of being unassociated with outcomes. The SRTR seeks to include in risk adjustment those variables whose estimated effects on the outcome, at least in terms of increasing versus decreasing survival, make sense from a clinical point of view.

Are there additional factors that we know or suspect to be clinically significant? Based on input from medical experts from the SRTR and the OPTN organ-specific committees, additional variables are tested for inclusion in the models. Some of these are only added to the models if there is strong evidence that they affect the outcome; others may be included regardless if the evidence is weaker but the common wisdom is that they are important.

Are data available? Is a comparison group for this practice or characteristic available or is the practice limited to just a few centers? As noted earlier, we can only adjust for factors that we know about all of the patients (or donors) in the country, and for types of operations performed at multiple transplant programs. There are many clinical features affecting patient survival, but if these factors are not collected uniformly and accurately, we cannot estimate the effect that those features 'usually' have on patients, and therefore cannot adjust for that effect.

Table 3 suggests the broad range of factors used in risk-adjusted analyses. More detail about the models, including lists of covariates, can be found in the technical documentation for the Program-Specific Reports at www.ustransplant.org.

The Role of Random Chance in Experiencing Good or Poor Outcomes

The 'true' ratio of observed to expected events that characterizes, or would result from, a particular program's treatment practices is not known. Discovering it would require that program to continue those exact practices on similar samples of patients forever, and would require us to observe those outcomes. The rates we observe and show in the Program-Specific Reports are a 'best estimate', based on recent performance. The underlying events, such as the death of a patient, are—like many daily phenomena—affected by random chance; these estimates, too, are affected by random chance.

Table 3: Risk-adjustment factor overview, 1-year graft survival

	Deceased donor kidney	Living donor kidney	Liver	Lung	Heart
Donor demographics					
Donor age	X	X	X	X	X
Donor race	X	X	X	X	
Donor comorbid history and other risk factors					
Donor cause of death	X		X	X	X
Donor comorbid history and factors	X			X	
Donor size			X	X	
Expanded criteria and DCD	X		X		
Recipient demographics					
Recipient age	X	X	X	X	X
Recipient insurance	X	X			
Recipient race	X	X	X	X	
Recipient sex				X	
Recipient diagnosis and functional status					
Diagnoses	X	X	X	X	X
Duration of illness	X	X			
Functional status indexes	X	X		X	
Physiologic reserve			X	X	X
Pretransplant treatment			X	X	X
Previous treatments	X	X	X	X	X
Recipient size	X	X			X
Recipient sensitivity					
Panel-reactive antibody	X	X			
Donor-recipient compatibility					
Blood type compatibility			X		
Donor relationship		X			
HLA mismatch	X	X			
Weight compatibility	X				
Organ transfer and travel characteristics					
Ischemic time	X			X	X
Kidney pumping	X				
Travel distance or sharing	X		X		

Source: SRTR, www.ustransplant.org. Note that this table is prepared as an example list, and is not exhaustive. Full listings of all covariates, separately for each organ, time period and type of survival (graft or patient) is available at www.ustransplant.org.

There is the inherent possibility that this 'best estimate' may be misleading, also because of random chance. It is important to determine the chances that the 'true' result—if we knew it—would suggest no systematic difference between the outcomes observed and expected, even when the best estimate suggests that there is a difference. Note that the process of risk adjustment does not affect the possibility that the differences we see between observed outcomes and expected outcomes occur because of random chance alone, which also exists without risk adjustment. In such cases, if the transplant program kept using the same practices and patient protocols forever, the observed difference between observed and expected outcomes would vanish because there is nothing systematically different about the program's practices, only a temporary streak of good (or bad) luck.

In Figure 4, line 10 shows the p-value or the probability that we would see such a difference (or more) between observed and expected outcomes, even if there was no systematic difference in treatment. Or to phrase it another way, the p-value describes the chance that the observed difference is due to chance alone, based on the sample size (number of transplants, patients, donors) and the size of difference observed. Given the same observed differences, a larger sample size implies a lower probability of being due to random chance; given the same sample size, a larger observed difference implies a lower probability that some difference is due to random chance.

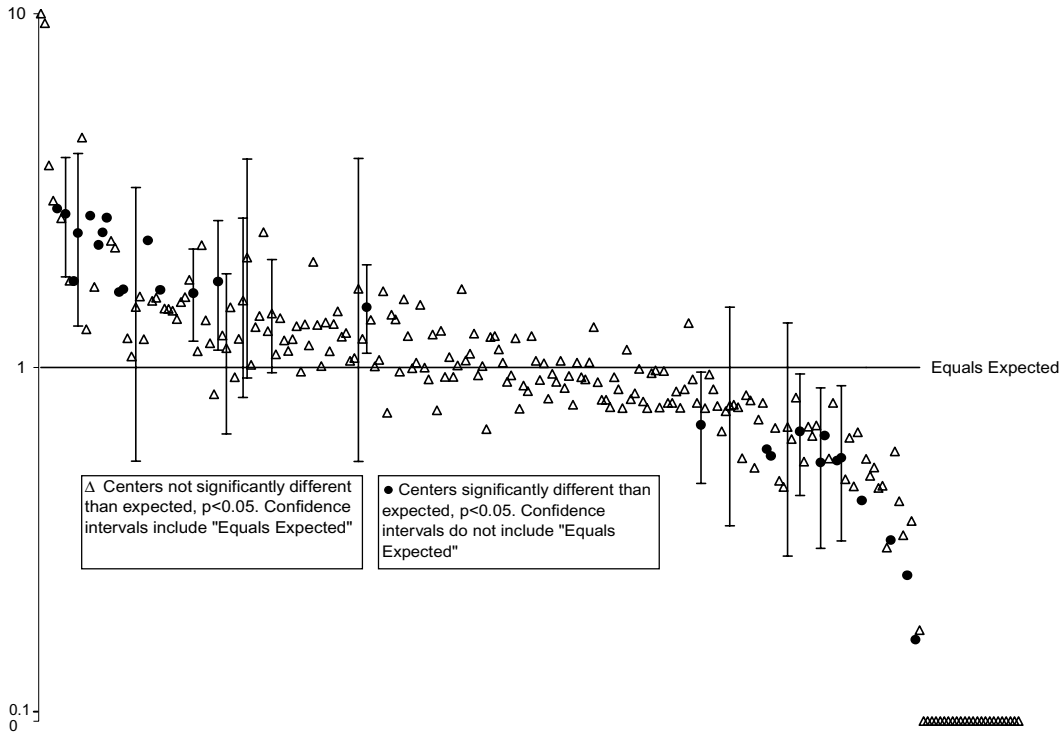
In the example case in Figure 4, the probability that the difference is as large or greater than observed due to random chance alone and not a systematic difference in treatment patterns is 0.469, or about 47%. In other words, even if this transplant program has no practices that would typically lead to poorer or better outcomes, there is nearly a 47% chance that we would see outcomes like this, or worse. So how unlikely, or how low a p-value, do we need in order to make conclusions about a program's performance?

When making conclusions regarding a hypothesis (e.g. classifying a program as under-performing), the level of 'statistical significance' depends on the acceptability of misclassifying a center as either under- or over-performing. For most statistical tests, a threshold of 5% probability (or a p-value of 0.05) of random chance is applied. This implies, for example, that a center with a ratio of observed deaths to expected deaths of greater than 1 would be classified as under-performing only if $p < 0.05$.

Note that the choice of a 5% probability threshold reflects the decision-maker's comfort with the possibility of misclassification. This comfort level might even depend upon the intended use of the classification system: if severe and automatic consequences result, a lower p-value threshold might be used; if mere 'further review' or examination of other factors results, a higher p-value threshold might be used. This level of comfort with misclassification is reflected in the principles implemented by the OPTN Membership and Professional Standards Committee, described below. For some purposes, statisticians use a more conservative threshold of 1% ($p < 0.01$). On the other hand, a p-value of 0.10, while not 'statistically significant' at this arbitrary level, still indicates that there is only a 10% chance that the observed discrepancy from expected outcomes was a random occurrence.

Another way to look at this random variation is the 95% confidence interval (CI), shown in line 9. If this center repeated its treatment practices on 100 different samples of patients with similar characteristics, the ratio of observed to expected outcomes would probably fall in this interval 95 times. The CI is intended to be a descriptor that reflects the precision of the estimate and, as a source of information, serves as a complement (rather than an alternative) to the p-value.

Figure 6 shows each program's adjusted graft failure ratio, indicating those that are significantly different from expected at a $p < 0.05$ level with a solid circle. While nearly all facilities exhibit a difference between observed and expected, only a small handful of them show differences unlikely (with less than 5% probability) to be considered random using this test. For illustration, vertical bars showing the confidence intervals are shown for several centers. Note that only when these confidence intervals do not overlap the 'equals expected' line, the differences are considered statistically significant.



Source: SRTR Transplant Program Specific Reports, published July 2007. Includes transplants performed in the 2.5 years ending 6/30/2006.

Figure 6: Some centers' standardized graft failure ratios are significantly different than expected; differences unlikely due to random chance.

Table 4: Audiences for Program-Specific Reports

Audiences	Purpose (and specific measures, if applicable)
Monitoring and process improvement	
HRSA / Division of Transplantation	• Identify problems with the organ transplantation system
OPTN Membership and Professional Standards Committee (MPSC)	• Identify individual transplant centers that may be under-performing or not following allocation policy • Upon further investigation of identified centers, review membership in the OPTN • *Measures: Posttransplant outcomes; others being explored (organ acceptance rates, transplant rates, etc.)*
Regulators and other payers	
CMS	• Review qualification for Medicare certification for both OPOs and transplant programs • *Measures: organs per donor by category, adjusted donation rates; posttransplant outcomes*
Private insurers	• Qualify transplant programs for preferred-provider plans • Identify individual transplant centers that may be under-performing, such that their customers (insured patients) are not well served • *Measures: Posttransplant outcomes; others as they appear on Standardized Request for Information (RFI)*
Others	
Media	• Identify and publicize problems either with the current system or individual centers, and help to explain the implications to the public
Transplant centers and clinicians	• Monitor performance in comparison to other centers • Be alerted before problems would arise to either monitoring or regulatory audience • Provide information to their patients about the performance of the center
Patients and families	• Learn about the performance of the transplant center caring for them • If choices exist, learn about the other centers • Find out about the general 'prognosis' for their disease

Source: SRTR.

Audiences and Use of These Statistics for Performance Measurement

The development and publication of these reports is a contractual requirement of the SRTR, for the intended benefit of the transplant system as a whole. In requiring that these statistics be publicly available, the Division of Transplantation (DOT) at the Health Resources and Services Administration (HRSA) recognizes that many different audiences—payers from the federal government and private sector, oversight committees from the OPTN, transplant professionals, the media and patients and their families—will use these data to contribute to the improved operation and outcomes from the transplant system.

Table 4 summarizes these audiences and their uses of these statistics.

The role of HRSA

Through the Department of Health and Human Services, HRSA is accountable to the Congress and the public for effective oversight of the nation's organ transplantation system. The DOT is ultimately accountable for monitoring and process improvement of the transplantation system.

HRSA's objectives are to achieve the best possible outcomes for transplant patients, by continuously improving knowledge about patient care and maximizing operational efficiencies. The DOT uses program-specific statistics to be apprised of issues facing the organ transplantation system; sometimes these issues are easier to see when viewed at a program-specific level. HRSA's oversight of the OPTN system is grounded in effective performance measurement using data submitted by OPTN members. Data—as found in the OPTN/SRTR Annual Reports and SRTR Program-Specific Reports—are used in multiple contexts, including monitoring the effectiveness of OPTN policies; generating short- and long-term outcome assessments; protecting the medical safety of patients undergoing solid organ transplantation; and, in the context of health informatics, to assess practice patterns and other risk factors to gain a better understanding of solid organ transplantation in the US.

Monitoring and process improvement

The Membership and Professional Standards Committee (MPSC) of the OPTN works to ensure that member transplant centers remain in compliance with criteria for OPTN membership. This role includes identifying centers that are not performing well, with the intention of helping them implement corrective action or reconsidering their membership. The performance measurement tools supplied by the SRTR to the MPSC help the committee identify transplant programs or OPOs that might require site visits or case reviews to look more deeply into potential problems.

Broadly speaking, the MPSC seeks to identify the most egregious of programs affecting the highest number of patients, including those with poor outcomes that:

(i) display a clinically significant pattern, suggesting a higher likelihood that practices contributing to poor outcomes might be identified, indicated by a high fraction of excess deaths (currently, at least 50% more deaths than expected, or O/E >1.5);
(ii) indicate that the magnitude of the problem, in terms of potential lives saved, should be sufficient to take action and place the center near the top of the priority list for action; currently, this requires that at least three more deaths are observed than were expected (i.e. O/E > 3);
(iii) are unlikely to be due to chance alone, with a probability of less than 5% (a one-sided p-value < 0.05).

Important in the implementation of these criteria, and resulting action, by the MPSC is the understanding of the possibility that random chance and incomplete adjustment (because of risk factors not collected in the data) affects these outcomes. Given that the MPSC uses these criteria to identify programs for further review, they are comfortable with a 5% probability that an identified program is misclassified. A one-sided p-value is used for this test because it quantifies the possibility that a result showing poorer outcomes, rather than poorer or better outcomes, is due to random chance. A two-sided test would be appropriate if, in addition to identifying under-performing centers, interest was also in identifying centers performing better than average (therefore the two-sided p-value and corresponding confidence interval are used on the public site).

These thresholds were set by the MPSC in an effort to identify the programs where reviews of practices might make the biggest differences in terms of the number of patients affected, while keeping within the resource constraints of the OPTN's review efforts (not selecting too many programs) and also minimizing the possibility of the burden of review on a center where the results may likely have occurred by random chance. Acknowledging that smaller centers are unlikely to be reviewed under these criteria, all smaller centers (performing nine or fewer transplants in a 2.5-year cohort) that experience at least one adverse event are subject to review. These thresholds are being reviewed by the SRTR and MPSC, as is the possibility of developing criteria that would incorporate the number of patients and events explicitly, hence precluding the need to classify centers as 'small.'

More information about exactly how these three principles are implemented, as well as their caveats, can be found in a previously published article on this subject (1).

Although these concepts are applied to posttransplant patient and graft survival at this time, the MPSC and SRTR are continually reviewing which additional outcomes may be useful in identifying programs for additional scrutiny. Currently, the MPSC is looking at different measures of waiting list outcomes (such as the waiting list transplant or mortality rate) and mechanisms that contribute to these outcomes (such as organ acceptance rate) to capture this phase of the process. Many of these outcomes are publicly available in the Program-Specific Reports.

Regulators and Payers

The Centers for Medicare and Medicaid Services (CMS) implemented similar concepts in their conditions of participation (COPs), which enable transplant programs to receive federal funding for transplant programs services. CMS has outlined a qualification system that is consistent with the thresholds chosen by the MPSC and described above.

CMS will use these outcomes data in conjunction with other data and information gathered from an onsite survey to measure a center's performance. The complete conditions of participation can be found at http://www.cms.hhs.gov/CFCsAndCoPs/downloads/trancenterreg2007.pdf.

Like CMS, many private insurers look to these performance measurements to certify transplant programs as providers of service. Many of these insurers participate in the standardized request for information (RFI) program, designed cooperatively with the OPTN Transplant Administrators Committee and the SRTR. These RFIs contain a wide range of statistics about transplant programs provided by the SRTR, many of them consistent with those found in the SRTR Program-Specific Reports. The focus of these reports, of course, is the section on posttransplant outcomes also found in Tables 10 and 11 of the transplant program reports. Several nonparticipating insurers also look to the public reports to find this information. Like the MPSC, many insurers use these rates as one of several indicators, often identifying programs where a closer look at patient protocols and outcomes is warranted.

Other users

Other private users, such as patients and families or the media, find the reports useful in identifying that there are, in fact, differences among transplant programs, and that those differences often involve trade-offs between performance in different phases of the process as described above.

For users who are interested in comparing transplant programs, the risk-adjusted ratio of observed to expected presents a clear advantage over ranking by observed survival rate (as survival rate may reflect either success or advantageous patient case mix). However, even ordering by this ratio is problematic because of differences in the variance of the estimated ratio among centers. Users

should remember that the p-values presented do not measure the statistical significance of the difference from other centers.

As described earlier, these statistics are not primarily intended to show whether a program is statistically significantly different than another program, only that the program is statistically different from expectations based on a national average that accounts for patient and donor characteristics. That is, while the reports may suggest that on any one measure, such as waiting list mortality, Program A has a higher than expected rate that is unlikely due to chance, and Program B does not, it is not true that differences between Programs A and B carry the same likelihood of being due to random chance. The statement that A is probably better than expected is a more valid statement than that A is probably better than B.

For this reason, the SRTR does not present a 'ranking' of centers, or ordered list by any of the measures, and discourages other users from doing so. Such a ranking might imply that there is a meaningful difference between any two adjacent centers (such as fifth vs. sixth on the list) when, in fact, statistical differences among such pairs is unlikely. This is particularly important for users who may wish to republish portions of these reports about their own center or a subset of centers.

Given These Performance Measures, Should Programs Avoid Difficult Cases?

Performance measurement metrics are intended to help identify facilities in which practices could be improved and to identify facilities that perform well, as well as to improve practices partly by understanding differences in achievement in different parts of the transplant process. The increased use of these metrics, most notably by CMS in their conditions of participation for federal funding, can produce mixed reaction from transplant programs and OPOs. In an effort to avoid scrutiny from the MPSC or CMS, some programs have suggested no longer treating difficult patients.

Very crude metrics such as target unadjusted survival rates or donation rates, if used by oversight agencies, regulatory or funding agencies or the general public, might logically dissuade programs from treating difficult patients, since it would lower these rates. However, as we have seen, the use of risk adjustment substantially diminishes the incentive to avoid difficult patients and donors. While adjustment cannot remove all disincentive for avoiding treating patients with measured and reported risks, it certainly goes well beyond the use of crude rates in removing this disincentive. And certainly, donors and patients with measured risk factors should not be avoided.

Should programs use Expanded Criteria Donors?

The risk-adjustment methods described earlier allow us to take into account many of the difficulties in treating specific kinds of patients. Earlier we described the mechanism by which transplant with an ECD resulted in a higher expected number of graft failures or deaths. Even so, some centers that are nearing the thresholds for review by CMS or MPSC shy away from accepting ECDs. Table 5 outlines the possible consequences of such decisions.

At the national level, in the top frame, the number of observed graft failures equals the number of expected graft failures, resulting in a standardized ratio of 1.00. This suggests that when the risk-adjustment model is applied, we get exactly the number of graft failures that we expect. Note that because we adjust for ECD in these risk adjustments (as well as all components of ECD determination), the number of graft failures equals expected for ECD and non-ECD (here called SCD, or Standard Criteria Donors) alike. The number of failures per transplant for ECDs, 0.15, is higher than that for SCDs, at 0.06: these organs do indeed have worse outcomes, but they are expected.

Both example centers shown also have worse outcomes for ECD organs than they do for SCD organs when measured in unadjusted failures per transplant. However, while Center A does not do as well as expected with its ECD organs (O/E = 1.17), Center B does better than expected (O/E = 0.84). Had Center B avoided performing these 16 ECD transplants, its standardized graft failure ratio would have risen from 1.32 (32% excess deaths) to 1.50 (50% excess deaths), even though these 16 had a higher number of failures per transplant.

Table 5: Effect of excluding ECD on program-specific kidney graft survival

	SCD	ECD	All Donors
All Centers			
Transplants	33918	4155	38073
Observed graft failures	2187	630	2817
Expected graft failures	2187	630	2817
Standardized ratio (O/E)	1.00	1.00	1.00
Failures per transplant	0.06	0.15	0.07
Center A			
Transplants	107	14	121
Observed graft failures	2	2	4
Expected graft failures	6.41	1.70	8.11
Standardized ratio (O/E)	0.31	1.17	0.49
Failures per transplant	0.02	0.14	0.03
Center B			
Transplants	102	16	118
Observed graft failures	9	2	11
Expected graft failures	5.98	2.37	8.36
Standardized ratio (O/E)	1.50	0.84	1.32
Failures per transplant	0.09	0.13	0.09

Source: SRTR Analyses of Program-Specific Reports released July 2007.

Center B is not uncommon: among all the programs performing ECD transplants in the July 2007 release of the Program-Specific Reports, about half (87 of 187) would have had a higher reported standardized ratio—indicating worse than expected outcomes—had they excluded ECD transplants.

Conclusions and Additional Information

In this article, we have addressed and synthesized the answers to many of the frequently asked questions about the Program-Specific Reports. Armed with this information, users of the Program-Specific Reports should understand:

(i) That these reports represent a complex solid organ transplantation system that often involves trade-offs between different types of outcomes.
(ii) That differences do exist among programs, and advanced statistical methods do a good job at helping us disentangle the effects of patient and donor characteristics from the performance attributed to the program or OPO. These methods and the data they use are imperfect and, therefore, only a very useful part of the evaluation process.
(iii) That these reports are not intended to be used to compare the performance of one program or OPO to another, but to identify programs whose practices may need to be corrected, avoided or emulated.

This description of the methods and concepts is intentionally left at an overview level. Further detail about these methods is available from the SRTR website, in the 'Background and Methodology' section of the Program-Specific Reports. Materials available there include a public-use slideshow covering many of these concepts and others used in the reports; technical journal articles describing these processes in more detail (3,4); and the technical documentation to the program-specific results, which includes current risk-adjustment model equations.

Acknowledgments

The Scientific Registry of Transplant Recipients is funded by contract number 234-2005-37009C from the Health Resources and Services Administration (HRSA), US Department of Health and Human Services. The views expressed herein are those of the authors and not necessarily those of the US Government. This is a US Government–sponsored work. There are no restrictions on its use.

This study was approved by HRSA's SRTR project officer. HRSA has determined that this study satisfies the criteria for the IRB exemption described in the 'Public Benefit and Service Program' provisions of 45 CFR 46.101(b)(5) and HRSA Circular 03.

This article was produced as part of the 2007 OPTN/SRTR Annual Report. The annual report gathers information on many aspects of solid organ transplantation in one publication. More information can be found at www.ustransplant.org.

References

1. Dickinson DM, Shearon TH, O'Keefe J et al. The 2005 SRTR report on the state of transplantation: SRTR center-specific reporting tools: Posttransplant outcomes. Am J Transplant 2006; 6: 1198–1211.
2. Cox DR. Regression models and life tables (with discussion). J Roy Stat Soc, Series B 1972: 197–220.
3. Levine GN, McCullough KP, Rodgers AM, Dickinson DM, Ashby VB, Schaubel DE. The 2005 SRTR report on the state of transplantation: Analytical methods and database design: Implications for transplant researchers, 2005. Am J Transplant 2006; 6: 1228–1242.
4. Schaubel DE, Dykstra DM, Murray S et al. SRTR report on the state of transplantation: Analytical approaches for transplant research, 2004. Am J Transplant 2005; (4 Pt 2): 950–957.