Developing Statistical Models to Assess Transplant Outcomes Using National Registries: The Process in the United States

Jon J. Snyder, PhD, MS,[1,2] Nicholas Salkowski, PhD,[1] S. Joseph Kim, MD, PhD,[3] David Zaun, MS,[1] Hui Xiong, MS,[1] Ajay K. Israni, MD, MS,[1,2,4] and Bertram L. Kasiske, MD[1,4]

Abstract: Created by the US National Organ Transplant Act in 1984, the Scientific Registry of Transplant Recipients (SRTR) is obligated to publicly report data on transplant program and organ procurement organization performance in the United States. These reports include risk-adjusted assessments of graft and patient survival, and programs performing worse or better than expected are identified. The SRTR currently maintains 43 risk adjustment models for assessing posttransplant patient and graft survival and, in collaboration with the SRTR Technical Advisory Committee, has developed and implemented a new systematic process for model evaluation and revision. Patient cohorts for the risk adjustment models are identified, and single-organ and multiorgan transplants are defined, then each risk adjustment model is developed following a prespecified set of steps. Model performance is assessed, the model is refit to a more recent cohort before each evaluation cycle, and then it is applied to the evaluation cohort. The field of solid organ transplantation is unique in the breadth of the standardized data that are collected. These data allow for quality assessment across all transplant providers in the United States. A standardized process of risk model development using data from national registries may enhance the field.

(*Transplantation* 2016;100: 288–294)

National registries of solid organ transplantation exist in several countries throughout the world. Many provide national reports on the numbers of candidates, donors, and transplants performed, whereas only a few are charged with monitoring patient outcomes and transplant program performance. In the United States, the Scientific Registry of Transplant Recipients (SRTR), as created by the National Organ Transplant Act in 1984[1] and codified by the Final Rule,[2] is obligated to publicly report data on transplant program and organ procurement organization performance. These reports include risk-adjusted assessments of graft and patient survival, and programs performing worse or better than expected are identified. The SRTR currently maintains 43 risk adjustment models for assessing posttransplant patient and graft survival (Table 1).[3] The statistical models are fit to the most recent data set during each 6-month reporting cycle. However, until now, no formal process for model evaluation and revision had been implemented. Therefore, SRTR began working closely with the SRTR Technical Advisory Committee (STAC) to develop a formal process for risk model development. This article describes SRTR's newly implemented process. Although the focus of this overview is on the development of first-year posttransplant patient and graft survival models, the concepts and processes can be generalized to other models, and SRTR is currently using this process to develop models to assess waitlist outcomes. The creation and dissemination of standardized processes for risk model development may aid other countries in developing or enhancing their national registries and facilitate international comparisons to improve the quality of transplant care and outcomes throughout the world.

Determining Patient Cohorts for Risk Adjustment Models

For each outcome assessed (ie, patient and graft survival), we define the cohort of patients for which the risk adjustment

SRTR risk adjustment models for posttransplant outcomes (as of December 2014)

| Transplant type | Risk adjustment models currently in use | | | |
| | Graft survival | | Patient survival | |
	1 y	3 y	1 y	3 y
Heart	A, P	A, P	A, P	A, P
Kidney (deceased donor)	A, P	A, P	A	A
Kidney (living donor)	A	A, P	A	A
Lung	A	A	A	A
Liver (deceased donor)	A, P	A, P	A, P	A, P
Liver (living donor)	A	A	A	A
Pancreas[a]	—	—	A	A
SPK	A[b]	A[b]	A	A
PAK	—	—	A	A

[a] Pancrease-alone transplants, excluding SPK and PAK transplants, which are modeled separately.
[b] Kidney graft survival only. Pancreas graft survival is not currently assessed.
A, adult model in use; P, pediatric model in use; PAK, pancreas after kidney; SPK, simultaneous pancreas-kidney.

models will be developed and to which the models will subsequently be applied. For example, it must be determined whether separate models should be developed for adult and pediatric patients, for recipients of living versus deceased donor organs, for single-organ versus multiorgan recipients, and for first versus subsequent transplants, and whether any patient populations should be excluded. The STAC has made a number of specific recommendations as to which populations should be evaluated (Table 2).

Defining Single-Organ and Multiorgan Transplants

The SRTR uses an algorithm to define single-organ and multiorgan transplants (Figure 1). First, a transplant event is reported to the Organ Procurement and Transplantation Network (OPTN) (box 1). A patient who receives more than 1 type of organ *from the same donor* is classified as a standard multiorgan recipient (box 3). Otherwise, the algorithm looks for another transplant record in the registry for the same recipient, occurring either before or after the current transplant. If the recipient underwent no other transplants, the transplant is classified as single-organ (box 8). If the

recipient underwent more than 1 transplant, the algorithm assesses whether any of the additional transplants met any of the following criteria:

- Transplant of 2 lung lobes from 2 living donors (box 5) is classified as a single-organ lung transplant.
- Transplant of a living donor kidney and a living donor liver on the same day (box 6) is classified as a multiorgan transplant.
- Transplant of a kidney from a living donor within 3 days of transplant of a pancreas from a deceased donor (box 7) is classified as a multiorgan transplant.

If none of these conditions are met, the event is classified as a single-organ transplant that was preceded or followed by another transplant (box 7 to box 8).

Single-Organ Models

Generally, for each single-organ transplant population, SRTR develops separate risk adjustment models for deceased donor and living donor recipients as applicable (Figure 2). For each of these 8 primary models, SRTR develops separate models for adult and pediatric recipients, and separate models for patient and graft survival. This yields 32 potential models for single-organ transplant recipients.

Risk Adjustment Model Development

The SRTR worked with the STAC to establish a standardized model-development process for all future risk-adjusted models (Figure 3).

Box 1: Cohort Construction

We first define the cohort used to develop the model. For example, to develop the model to assess first-year graft survival of adult single-organ kidney transplant recipients (Figure 2, model 4), we used patients who underwent transplant during the most recent 2.5-year period ending on either June 30 or December 31. This ensures that at least 6 months of follow-up is available for all patients in the cohort.

Box 2: Requirement for at Least 25 Events in the Development Cohort

To develop a valid risk adjustment model, an adequate number of events (in this case graft failures or deaths) must

Cohorts for which outcomes are monitored in separate risk-adjusted models

| Recipient population	Recommendation
Adult (aged ≥ 18 y) versus pediatric (aged < 18 y) recipients	Adult and pediatric recipients should be analyzed separately with separate risk adjustment models. Adult models are applied to adult recipients and pediatric models to pediatric recipients regardless of whether the transplant took place at a facility specializing in pediatric transplant.
Deceased versus living donor recipients	Recipients of deceased and living donor organs should be analyzed separately with risk adjustment models developed specifically within those populations.
Recipients of first versus subsequent transplants	Assessment of graft outcomes should include first and subsequent transplant recipients with appropriate adjustments made for retransplants. Analysis of patient survival is limited to first-transplant recipients such that deaths are counted only once.
Single-organ versus multiorgan recipients	Risk adjustment models are currently developed on single-organ recipients only except for simultaneous kidney-pancreas and simultaneous heart-lung recipients. SRTR is developing models for other common multiorgan transplants (eg, simultaneous kidney-liver).
US citizen versus non-US citizen	The US registry contains data on all recipients who underwent transplant in a US transplant program, and policies regarding patient follow-up are the same regardless of citizenship or place of permanent residence. Therefore, all recipients are included as long as the transplant took place at a US transplant program.
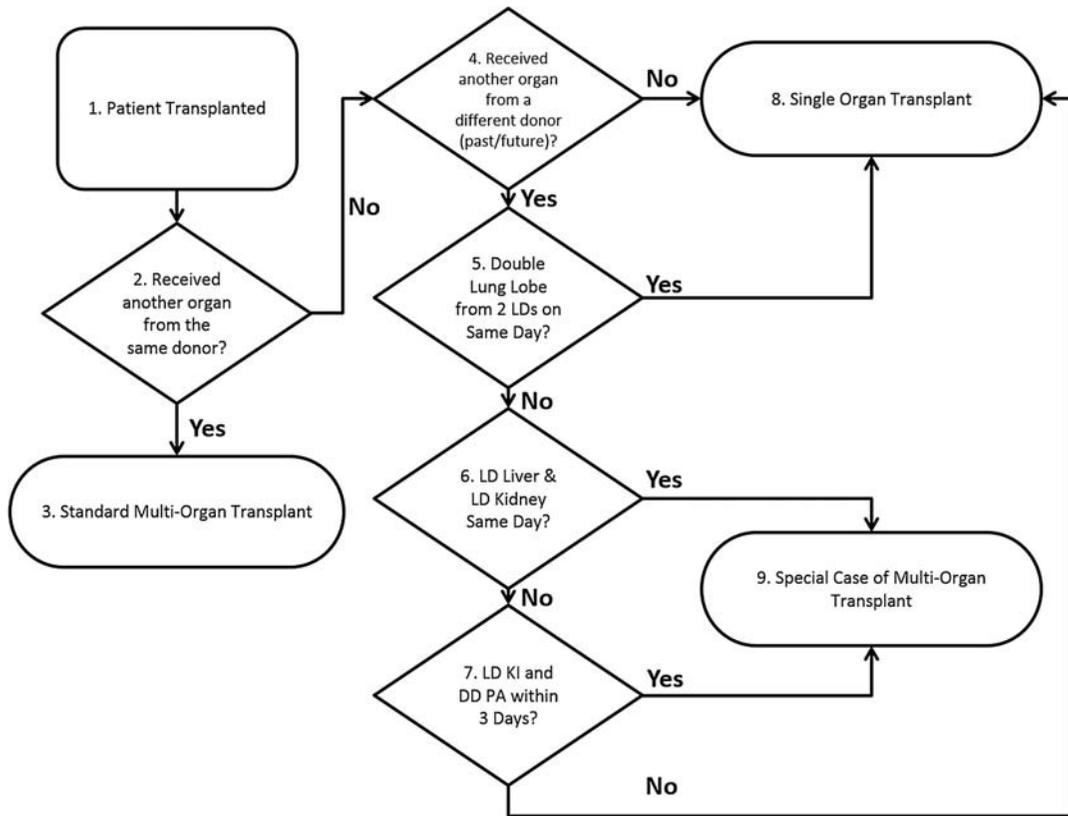
FIGURE 1. Algorithm defining single-organ and multiorgan transplants. DD, deceased donor; KI, kidney; LD, living donor; PA pancreas.

have occurred in the development cohort to allow selection of variables to adjust for risk. A proposed convention in multivariate prognostic modeling is to require at least 10 events per risk adjuster in the final model.[4] The STAC recommended a more conservative minimum of 25 events in the development data set to attempt to build a risk adjustment model (approximately 10 events per year over the 2.5-year period). The more conservative approach was recommended to avoid attempting to derive expected outcomes in a population with very few events observed overall. For example, in 3 of the pediatric kidney recipient populations assessed, fewer than 25 events occurred nationally. When fewer than 25 events are observed, SRTR will not attempt to derive a risk adjustment model and will not calculate expected outcomes (Figure 3, box 3). If regulators wish to monitor outcomes in populations with few events, we believe a "safety net" approach may serve the purpose, for example, reviewing a program with less than 75% success rate and more than 1 failure, where the exact thresholds could be determined based on the national failure rate for the organ type being evaluated.

Box 4: Data Preparation

In addition to defining the cohort to be included (described above), we also must decide which variables should be considered for risk adjustment. The SRTR uses a process to solicit expert opinion from OPTN organ-specific committees, whose members are experts in the particular transplant type. The SRTR compiles reports detailing key data elements collected by OPTN and, in conjunction with the committee, identifies a set of variables that are considered potentially appropriate for risk adjustment. Variables that occur after transplantation (eg, need for dialysis within the first week

for kidney recipients) or directly reflect patient care options (eg, use of specific immunosuppressive medications) are not included for consideration.
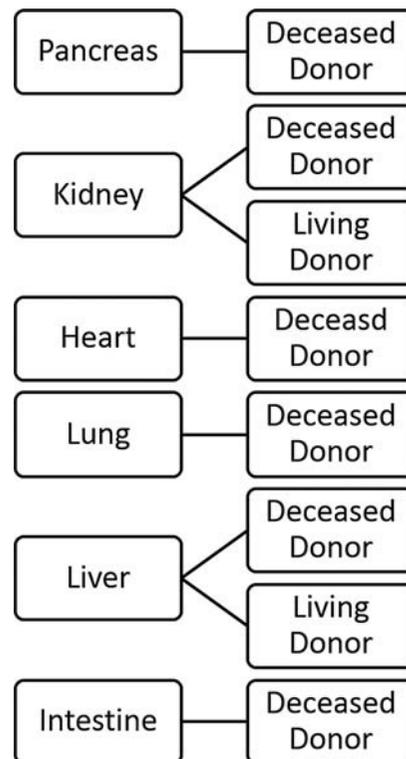

FIGURE 2. Single-organ transplant risk adjustment models to be considered.
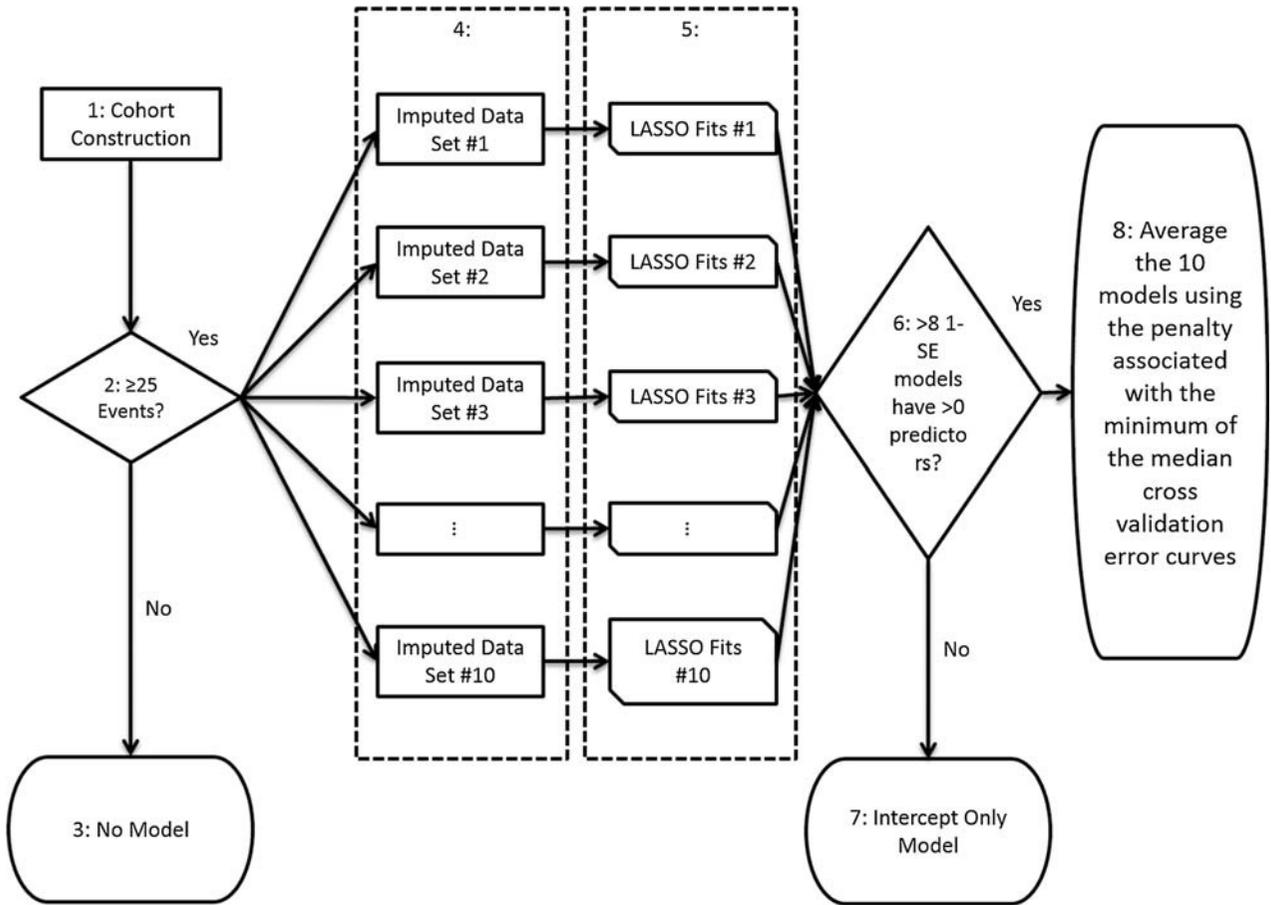
FIGURE 3. Process for risk adjustment model development.

A broad range of variables is included in the report to the organ-specific committee, and the committee can recommend adding or deleting specific variables. The report contains information on the distribution of data elements, the frequency of missing data, the frequency of outliers (extreme values), and the relationship between the data element and
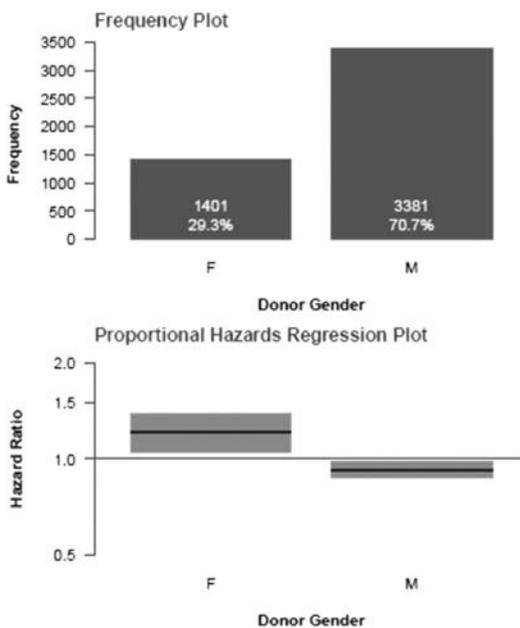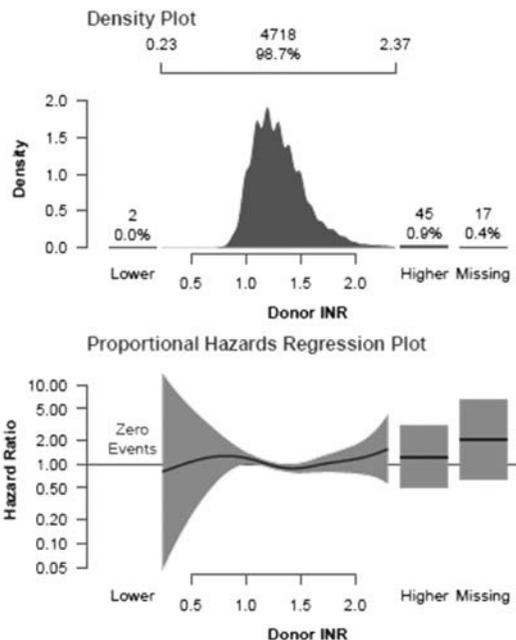
FIGURE 4. Examples of variable reports for expert review. Shown are donor gender (left panel) and donor INR (right panel) created for the adult heart transplant first-year graft failure model. INR indicates international normalized ratio.

the outcome of interest. Committee members are asked to provide their opinions about the appropriateness of including the element and the importance of the element for stratifying recipient risk. Figure 4 provides an example of a report showing the distribution of donor gender (left panel) and donor international normalized ratio (right panel). The unadjusted relationship between donor sex and graft failure is shown in the bottom left panel. The unadjusted relationship between donor international normalized ratio and graft failure is then provided as estimated from a Cox proportional hazards model using a penalized smoothing spline parameterization within the typical range of the variable. In addition, the hazard ratios for missing and outlier values (bottom-right panel) are estimated.

For each element, committee members are asked to answer the following questions:

(1) Should the element be included in the risk adjustment models?
(2) If not, why not? Choices include (multiple selections possible):
  (a) Clinically irrelevant: you do not believe this element has any relationship with outcomes.
  (b) Data are unreliable: you do not believe this element, as currently collected, is reliable. Reasons might include nonuniform understanding of definition or nonuniform time at which the element is collected, for example, how long pretransplant a certain laboratory value was measured.
  (c) "Gameable": You believe a program could attempt to record a level for this element that affords it the best risk adjustment.
  (d) Inappropriate for SRTR risk adjustment: some elements that could be adjusted for might be deemed inappropriate for inclusion in the risk adjustment models, such as elements that could be considered clinical care or a clinical decision.
  (e) Other: indicate if there are other reasons we should not consider adjusting for the element.

In asking these questions, we attempt to understand if the committee members think the element is important to include, either because it is related to outcomes or because it increases the face validity of the models, ie, if the community would have greater trust in the models if this element were included. Finally, we ask the committee if we missed any elements in the original compilation of the report.

Following the committee's review, SRTR staff may alter the model development data set by adding, removing, or altering variables. The SRTR further refines categorical elements by combining levels to ensure adequate numbers in each category.

Box 5: Imputing Missing Values

Once the data set is constructed, we next deal with missing or "unknown" responses to the data elements. Historically, SRTR risk adjustment models have handled missing/unknown responses by modeling these effects separately, that is, a separate adjustment for "missing" or "unknown." This is not ideal for the purpose of risk adjustment because missingness may (1) bias estimates of the nonmissing levels of the variable if missing data are not completely random[5] and (2) yield a perverse incentive for programs to leave a known value missing or unknown because doing so would impart better risk adjustment. Modeling missing or unknown values as a distinct level has historically allowed programs to benefit from

leaving a variable missing or unknown, even when the variable is missing completely at random. However, to avoid this bias, SRTR now develops risk adjustment models without separate risk estimates for missing values. Currently, a value can be "missing" in the following situations:

(1) A value is not entered for the variable (*Note*: as of March 31, 2015, OPTN removed all nonrequired elements from data collection forms; thus, in the future SRTR will not consider any non-required fields during model development).
(2) *Unknown* or an equivalent response is entered. Many elements on current OPTN data forms allow for an *Unknown* response (eg, "Working for income?" can be answered *Yes*, *No*, or *Unknown*). SRTR considers *Unknown* responses to be missing.
(3) A value that is outside the plausible range of values is entered. The reported value will be treated as missing. The "plausible range" for continuous variables (total albumin, height, weight, and so on), is determined by expert opinion (SRTR staff, STAC members, and OPTN committee input) and is generally conservative in allowing extreme but potentially plausible values.

By these definitions, in the new kidney models developed, at least 1 value was missing in approximately one-third of candidate variables, and up to 14% of some variables were missing and imputed. The SRTR implements an imputation process to fill in missing values with modeled estimates using multiple imputation by chained equations.[6] The imputation routine uses predictive mean matching for continuous variables, logistic regression for binary variables, and multinomial logistic regression for categorical variables. The imputation routine results in 10 multiply imputed data sets, each with distinct modeled estimates of the missing values. Each of these 10 data sets is then used to build 10 distinct risk models in the next step of the process.

Box 6: Variable Selection

The above steps yield 10 model development data sets without missing values. These data sets include many variables for potential inclusion in the final risk adjustment model, often more than 100; however, developing the smallest model that achieves the goal of providing the best risk adjustment is the ultimate goal. SRTR could include all available data elements in the risk adjustment model. This is problematic because (1) it results in statistical overfitting, and many variables have no meaningful effect on risk adjustment. 2) It increases the burden on programs to ensure the accuracy of their data. The SRTR provides each program with its patient data that will be evaluated in each 6-month reporting cycle, allowing programs to review their data to ensure its accuracy. We prefer to provide programs with fewer elements for review, excluding the many elements with no substantial impact on risk adjustment.

The SRTR uses the least absolute shrinkage and selection operator (LASSO) for Cox proportional hazards models to arrive at a parsimonious model.[7] The LASSO is a method for choosing the most predictive set of variables from a larger set of possible predictors. Before the LASSO is run, the input data sets are modified to truncate values for continuous variables at the first and 99th percentiles (values outside these limits are set to the first or 99th percentile value as appropriate), so splines are not estimated in the extreme tails of the distribution. Categorical variables are overparameterized to

include a dummy-coded predictor for each level of the categorical variable (yielding N dummy variables rather than N-1). Linear splines are explored to represent continuous variables following recommendation 1.8 from a consensus conference convened by the US Health Resources and Services Administration, OPTN, and SRTR in 2012: "Avoid converting continuous data elements to categorical elements, and use smoothed splines only when continuous linear values are not appropriate.[8]" Recognizing that a linear trend may not capture a complex nonlinear relationship, the routine allows for knots (or bends in the line) at any of 9 points as determined by the deciles of the distribution. This allows for linear approximations to complex nonlinear relationships. Exploratory analyses were performed to compare the performance of linear splines with natural cubic splines that may better approximate nonlinear relationships. Linear splines yielded comparable performance and were deemed easier to communicate, so we pursue linear spline parameterizations of continuous variables, allowing up to 9 bends in the line to best fit the data. Categorical variables are overparameterized to allow the LASSO procedure to choose the optimal reference level(s). The vast number of potential predictors makes assessment of all interaction terms impractical. Therefore, model development focuses on main effects unless there is a compelling reason to explore an interaction term. Clinical expertise, previous literature, and recommendations of the organ-specific OPTN committees are considered when choosing which interactions to explore, for example, recipient-donor height ratio for lung recipients.

The LASSO procedure is run on each of the 10 multiply imputed data sets. Ten-fold cross-validation is used to determine the cross-validation error at each level of the constraint parameter. Before deciding on a final model, we pause at this step to determine if any of the multiply imputed data sets yield at least one good risk predictor according to the LASSO procedure.

Boxes 7 and 8: Does the LASSO Procedure Indicate at Least 1 Good Risk Predictor?

Once the LASSO procedure is run on each of the 10 multiply imputed data sets, we examine the models associated with the minimum cross-validation error that is within 1 standard deviation of the estimated overall minimum cross-validation error for that data set. This model will include fewer predictors than the model yielding the minimum cross-validation error, but its prediction capabilities will be reasonably consistent. The LASSO may indicate that none of the available variables yield substantial predictive capability; therefore, each of these 10 models may or may not contain risk predictors. If at least 9 of the 10 models do not contain at least one risk predictor, we stop and conclude that none of the available data improve our ability to predict risk in the recipient population. In this case, we simply use the observed national rate of failures to derive estimates of expected survival in the population (box 8). If 9 of 10 data sets indicate at least 1 risk predictor, we continue to build the final risk adjustment model.

Box 9: Deriving the Final Risk Adjustment Model

Having at least 1 good risk predictor, we combine results from the 10 multiply imputed data sets to arrive at a final

model. For each level of the constraint parameter, the median generalized cross-validation statistics are calculated across the 10 multiply imputed data sets. The final LASSO constraint parameter is then determined to be the one that yields the minimum of the median cross-validation errors across the 10 data sets. Each of the 10 LASSO-fit models at this constraint parameter is kept, and the final parameter estimates are calculated as the mean of the parameter estimates across the 10 sets. If a risk predictor was kept in some but not all of the 10 LASSO fits, the parameter is assumed to be 0 for the models that did not include the predictor. This process yields the final risk adjustment model.

Model Performance and Validation

Once the described process produces final models, model performance is assessed. The model's ability to discriminate low-risk from high-risk transplants is assessed using the C (or concordance) statistic adapted to the setting of the Cox proportional hazards model. The model's ability to accurately predict the outcome is assessed through use of calibration plots of expected versus observed event counts within deciles of predicted risk obtained from the models. The STAC and OPTN's organ-specific committees review the models to assess face validity. Construct validity (ie, whether we are measuring what we think we are measuring) is assessed by demonstrating separation of relevant outcomes by predicted risk group through the calibration plots and Kaplan-Meier plots stratified by deciles of predicted risk. At the time of this report, the new process has been applied to the development of models for kidney and heart recipients. The kidney models achieved C statistics ranging from 0.66 to 0.76 and the heart models from 0.67 to 0.83. In some cases, C statistics improved by 25% over previous models. Finally, SRTR does not report pseudo R-squared values for the models, given the many variations available for pseudo R-squared statistics in the setting of Cox proportional hazards models, each with strengths and weaknesses, making interpretation difficult.

Risk Adjustment Model Refitting

Once the final model has been determined, SRTR refits it to a more recent cohort before each evaluation cycle, so the model coefficients can be updated during each cycle before the next full model rebuild performed after 3 years. To accomplish this, SRTR uses the current program-specific report evaluation cohort and refits the models using the process described above; however, the LASSO penalty is predetermined to be the one used during the initial model development phase. This produces a model with the same set of risk predictors as determined during the initial model development phase, but allows the model coefficients to adjust. Thus, the structure of the models remains the same (includes the same set of predictors), but the model coefficients adjust to the more recent data.

Risk Adjustment Model Application

Once the final risk adjustment model has been developed and refit to a recent cohort, it must be applied to the cohort of patients to be evaluated, that is, the evaluation cohort. The evaluation cohort consists of the most recent 2.5-year cohort of transplant recipients, defined similarly to the model development cohort: the most recent 2.5-year period ending on either June 30 or December 31, 6 months before the date of evaluation. For example, for evaluations commencing in

October 2014, the cohort for first-year outcomes evaluations consists of patients undergoing transplants between July 1, 2011, and December 31, 2013.

Handling of Missing Values During Program Evaluation

Risk adjustment models developed as described do not include separate risk adjustment for missing values (with missing as defined above). Therefore, if data are missing during the evaluation process, we must assign a level of risk. In response to a recommendation from the consensus conference to "substitute missing data with values that are least favorable to the center, thus encouraging centers to accurately record data,"[8] the STAC recommended that, in the absence of reported data showing that the patient was riskier than the lowest risk level, SRTR will assign the lowest risk for that characteristic. For continuous predictors, a risk level equivalent to the lowest risk level within the plausible range of values would be assigned. As described previously, linear splines are developed between the first and 99th percentiles of the data. Reported values outside of this range, but within the plausible range, are assigned risk consistent with the first or 99th percentile as appropriate. The plausible range is decided before model development. Programs can review their data before the final evaluation and correct any values identified as missing, unknown, or otherwise outside of the plausible range. If, during program evaluation, a data point remains outside of the plausible range, SRTR assumes the value is missing and assigns the lowest risk for that element.

Once missing values have been assigned, the risk adjustment models are used to derive expected event counts for each individual in the evaluation cohort. Patient level observed and expected event counts are then summed to arrive at program level observed and expected event counts as reported in the program-specific reports.

Risk Adjustment Model Maintenance and Process Oversight

Given the number of risk adjustment models SRTR develops and maintains, rebuilding each model during each 6-month evaluation cycle is not possible. Therefore, SRTR has implemented a revolving cycle of model rebuilding in the following order: kidney, heart, lung, liver, intestine, and pancreas.

Although the models will only be completely rebuilt based on this cycle, the STAC recommended that models be *recalibrated* during each 6-month evaluation cycle by refitting the current model (using the same risk predictors and parameterizations) to an updated but lagged cohort of patients, as described previously. The STAC will continue to provide process oversight.

Summary

The field of solid organ transplantation is unique in the breadth of standardized data collected through OPTN in the United States and similar registries in other nations. The SRTR is charged with providing program performance metrics publicly and to the OPTN's membership and Professional Standards Committee to assist in system quality monitoring in the United States. To enhance credibility and trust, it is imperative that the process used to develop the models be transparent. The STAC will continue to provide oversight and guidance to improve the statistical processes that SRTR uses. However, the process is entirely dependent on accurate and relevant data collected by OPTN. OPTN's newly formed Data Advisory Committee is charged with reviewing the OPTN data collection system, recognizing its key role in risk adjustment for the purpose of program evaluation. These processes will continue to enhance the system used to evaluate transplant program performance. A similar approach to the systematic development of risk adjustment models among the statistical entities charged with reporting quality data in their respective countries could help ensure more comparable statistics for transplant outcomes, ultimately driving quality improvements in the field at the global level.

REFERENCES

1. Department of Health and Human Services. National Organ Transplant Act of 1984. Public Law 98–507, Title 42-Section 273.10-19-0984. 98 Stat. 2339–2348. Available at: http://www.gpo.gov/fdsys/pkg/STATUTE-98/pdf/STATUTE-98-Pg2339.pdf. Accessed April 15, 2015.
2. Department of Health and Human Services. Public Health Service, Final Rule. 2000. 2015. Available at: http://www.ecfr.gov/cgi/t/text/text-idx?c=ecfr&tpl=/ecfrbrowse/Title42/42cfr121_main_02.tpl. Accessed April 15, 2015.
3. Scientific Registry of Transplant Recipients. Risk-adjustment models. Post-transplant baseline survival curves (all organs). 2015. Available at: http://www.srtr.org/csr/current/modtabs.aspx. Accessed April 15, 2015.
4. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361–387.
5. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581–592.
6. van Burren S, Groothuis-Oudshoorn K, Robitzsch A, et al. Mice: Multivariate imputation by chained equations; 2015. Available at: http://cran.r-project.org/web/packages/mice/index.html. Accessed April 15, 2015.
7. Tibshirani R. The LASSO method for variable selection in the Cox model. *Stat Med*. 1997;16:385–395.
8. 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.