SRC-AMS Meeting Minutes

Analytical Methods Subcommittee Teleconference

August 21, 2023, 1:00 PM – 3:30 PM CDT

Voting Members:
David Vock, PhD (Co-chair)
Shu-Xia Li, PhD
Brent Logan, PhD
Katherine Panageas, PhD
Erika Helgeson, PhD
William (Bill) Irish, PhD
Megan Neely, PhD

Not in Attendance:
Andrew Schaefer, PhD

Ex-Officio Members:
Jon Snyder, PhD (Co-chair)

HRSA:
Adriana Martinez, MS

SRTR Staff:
Ryutaro Hirose, MD
Larry Hunsicker, MD
Ajay Israni, MD, MS
Grace Lyden, PhD
Jon Miller, PhD
Josh Pyke, PhD
Tim Weaver, MS
Nicholas Wood, PhD
David Zaun, MS

Welcome and opening remarks

Dr. Jon Snyder and Dr. David Vock called the Analytical Methods Subcommittee (AMS) meeting to order. Dr. Snyder reviewed the agenda and conflict of interest management and then proceeded with the first item.

AMS membership & nominating process

Dr. Snyder reviewed the new nominating process for the SRTR Review Committee (SRC) and three subcommittees, including AMS. The process was created for transparency in choosing new members and the opportunity to be considered as a candidate. SRTR is trying to achieve 3-year terms with approximately one-third of the committee rotating off each year. Dr. Shu-Xia Li and Dr. Katherine Panageas will finish their terms at the end of December 2023.

Nominees will be reviewed by the Nominating Committee, with recommendations brought before the SRC for final recommendations to SRTR leadership. The application process will open the first week of September and close October 6, 2023. Application materials will be posted to SRTR.org. Final recommendations will be made at the SRC meeting on October 27, 2023. Dr. Vock emphasized that members of the AMS do not necessarily need to have experience in the field of transplantation.

CMS's organ procurement organization performance metrics

Dr. Snyder overviewed the Centers for Medicare & Medicaid Services (CMS) metrics for evaluating the performance of organ procurement organizations (OPOs) and the 3-tier system for deciding recertification status. CMS will be making recertification decisions for OPOs in 2026 based on data

from calendar year 2024. SRTR believes the chosen boundaries used to determine the OPO tier are biased against larger OPOs and sought the committee's feedback on the SRTR analysis and explanation of this characteristic of the CMS evaluation system. Dr. Snyder explained why SRTR believes the targets, based on the 75th and 50th percentiles of all OPOs' performance in the prior year, are biased against larger OPOs, and sought feedback from the members on the statistical analysis and explanation.

Dr. Snyder first gave some background information on the 56 OPOs in the United States. Each OPO serves a geographic region of the country, and they are set up by federal contract with CMS to manage the deceased donor organ procurement process within that geography. Dr. Ryutaro Hirose added that each OPO varies greatly in terms of geographic landmass, population, and demographics.

Two metrics were published by CMS in December 2020: the donation rate and the transplant rate. The numerator for the donation rate is the number of donors from whom at least one organ was transplanted or the pancreas was sent for research. The denominator is from the Centers for Disease Control and Prevention's (CDC's) detailed mortality data and is calculated using the "cause, age, and location consistent" (CALC) deaths methodology based on a series of *International Classification of Diseases, 10th Revision (ICD-10)* codes in the primary cause of death on the death certificate. Risk adjustment is not applied. For transplant rate, the denominator is the same as the donation rate denominator. The numerator counts how many organs were transplanted, including pancreata sent for research. The transplant rate is adjusted for the decedent's age.

Dr. Snyder reviewed the 3-tier system CMS applies to each metric. An OPO is assigned to tier 1 if their metric is not significantly below the 75th percentile of all OPO performance in the prior year. Tier 1 OPOs are recertified for an additional 4 years. OPOs are assigned to tier 2 if their metric is significantly below the 75th percentile but not the median of the prior year. These OPOs have the opportunity to recompete for their contract, and other OPOs can also bid for it. Tier 3 OPOs are significantly below the median of the prior year and are decertified.

Dr. Snyder showed data from the current evaluations of OPOs. Dr. Larry Hunsicker pointed out that the lower the number of potential donors in the OPO's service area, the less precisely the donation rate or transplant rate will be estimated, resulting in a wider distribution of estimates—meaning there is a better chance for small OPOs to have a high or low rate just by chance. Dr. Snyder showed box plots representing the distribution of OPO size by tier for donation and transplant rates. He noted that the size distribution for tier-1 OPOs is shifted lower for both rates compared to tiers 2 and 3.

Dr. Snyder shared graphs of simulated donation rates by OPO volume (assuming all OPOs perform exactly the same at the national average of 11.4) to help explain these concepts to the transplant community. He simulated about 500,000 OPOs across the range of sizes and plotted the results. He further explained that the 75th percentile of the distribution would be higher for smaller OPOs and lower for large OPOs; however, CMS establishes the 75th percentile without taking OPO size into consideration. Therefore, the national 75th percentile is too low for small OPOs and too high for large OPOs, making it easier for small OPOs to achieve the target just by chance alone.

Dr. Grace Lyden explained a simulation she ran from the perspective of the confidence interval, type I error rates, and statistical power. Confidence intervals are very good at preserving the type I error rate across samples of different sizes, but are not guaranteed to balance the type II error rate (1 minus power). Dr. Lyden explained that, for smaller OPOs, it is easier to not reject the null hypothesis that the OPO is not below the 75th percentile.

Dr. Snyder showed figures that illustrate the curved nature of the flagging boundaries used for transplant program evaluations, which differ by transplant program size, in contrast to the boundaries chosen by CMS, which are constant relative to OPO size. He also mentioned an SRTR article published in the *American Journal of Transplantation* in 2020 that noted the boundaries chosen by CMS would be biased against larger OPOs.

Dr. Snyder asked members if they agreed with this assessment, and if there are ways to better explain this to the transplant community. Dr. Vock thought the graph Dr. Lyden presented was the most persuasive. Dr. William Irish agreed but thought the other graphs would resonate better with a nonstatistical audience. A median test had not been done on Dr. Lyden's simulation, but it was suspected the same problem would occur for the tier-2/tier-3 boundary at the median; although Dr. Snyder had hypothesized that the problem would not affect this boundary. Dr. Brent Logan and Dr. Megan Neely asked if the curved boundary fully corrected the problem. Dr. Neely suggested exploring solutions to the problem since this has not been done.

Dr. Logan also questioned if a curved boundary would have a major impact on correcting the issue. Dr. Snyder pointed out that in the posttransplant outcomes flag zone, there was no large effect across volume for transplant programs. The shrinkage applied from the Bayesian estimators on the transplant program side was meant to address this issue, and requiring a certain level of probability that programs are above this target boundary results in the curved nature of the boundary. However, Dr. Logan pointed out that the shrinkage estimator will have the same power issues, with the small centers shrunk back substantially. Dr. Vock suggested generating a donation rate for each center that is variable but does not vary based off of the CALC death to see if the phenomenon still exists. Dr. Hirose said there may be other reasons (in addition to the ones already discussed) why the CMS assessments are biased, such as lack of risk adjustment for cause of death or other factors that may influence the likelihood of donation. SRTR agreed to do more exploration and extend the simulation to address the committee's recommendations.

Long-term outcomes app

Dr. Jon Miller went over the long-term transplant outcomes application (app) that is under development. The app was a recommendation from the Task 5 consensus conference. A summer intern developed a new app for patients to look at up to 10-year outcomes posttransplant. The estimates were derived from Cox proportional hazard models with specific variables relevant for patients. It began with heart transplants and built on previous period prevalent methodologies. The initial cohort was roughly 25,000 adult and 4,000 pediatric transplant recipients, from January 2013 through January 2023.

Dr. Miller presented a static screenshot of the app, which included a slider bar at the top for predicting outcomes up to 10 years. The app currently presents an estimated survival curve responsive to the selected patient characteristics and requested follow-up time to predict. To offer

patients some ability to engage with potential decisions in the transplant journey, the variable of donation after cardiac death (DCD) was added along with donation after brain death (DBD). SRTR is working to include the ability to plot the two curves separately. Dr. Miller said SRTR was looking to implement heart, kidney, liver, and lungs into the app. There are also discussions of how the model building should be supervised and automated, and whether tree-based machine learning methods should be explored.

Dr. Erika Helgeson asked if cross validation was used when building the model. Dr. Miller said that while this was not done for initial model development, it is typically a part of SRTR workflow and will be addressed as the model continues to be developed. Dr. Lyden suggested a test of the predictions based on historical development cohorts, which would be a good way to look at period prevalent versus incident with time effect to see which of those yields the most accurate predictions. Dr. Helgeson agreed with this idea of having different periods for testing. Dr. Helgeson added that patients might want to know how much variability exists in the estimated survival curves. Dr. Miller said confidence intervals could be explored. Dr. Hunsicker said he was against trying to fit confidence intervals because they would be the confidence interval of the estimate of the fraction of people surviving over periods of time. He said they would be interpreted as what is the chance a patient is going to survive, which is not true. He thought confidence intervals would be misleading and hard to understand. Dr. Hirose said this is a problem many clinicians have in explaining these terms to patients.

Dr. Hirose said with period prevalence work, it can be difficult to explain to the patient that the period prevalent cohort is highly selective and excludes all the patients who died, and includes those who did not die in the first 2 to 5 years. He asked if this could be easily explained to a patient. that looking at period prevalent data as opposed to following the entire cohort out for that period of time. Dr. Hunsicker clarified that the period prevalent approach yields a conditional estimate of survival at each point along the survival curve, similar to how incident-based survival curves are constructed.

Dr. Lyden asked members if they had experience using off-the-shelf methods or machine learning methods for survival outcomes, and scaling modeling across a number of outcomes and cohorts. Dr. Logan said he found from his experience with bone marrow transplant and stem cell therapy that machine learning approaches are fairly data intensive, yield only modest benefits, and require a lot of tuning. Dr. Logan also said that having pediatric and adult outcomes together gives a slightly more efficient estimation. He suggested trying this method to see what the out-of-performance sample ends up looking like. Dr. Ajay Israni added that from a clinical standpoint for kidneys, stratification by deceased versus living donor transplants should be done. Dr. Lyden confirmed that this is being done.

SRTR will continue development of the long-term outcomes app and Dr. Snyder thanked the committee members for their insights.

Allocation policy simulation

Dr. Snyder informed the subcommittee that one of SRTR's roles is simulating organ allocation policy. SRTR presents many metrics back to the Organ Procurement and Transplantation Network (OPTN) committee members showing the results of the simulations. These metrics are currently unadjusted,

and SRTR believes there are reasons to present adjusted analysis when subgroups of patients are presented. For example, when presenting transplant rates within blood types, SRTR would like to adjust for other factors that may influence allocation priority if those differ by blood type. SRTR is seeking the committee's feedback on this approach.

Mr. Tim Weaver first reviewed the framework for SRTR simulation of the organ allocation system and how the simulation is used to estimate event rates and other summary measures for each allocation system being evaluated. He reviewed an example from a recent simulation study with four different panels for each blood type (A, AB, B, and O). The example showed transplant rates for different allocation policies. The high-low bars represent a range across the stochastic iterations, or the mean, minimum, and maximum values across simulate scenarios.

Dr. Snyder reiterated the concern that simply presenting unadjusted rates may not be sufficient if there exists confounding by other factors that may impact the candidate's allocation priority. Mr. Weaver reviewed options for adjusting rates or otherwise dealing with potential confounding. The first was further subsetting to create smaller groups by potential confounders. The second was to calculate standardized rates based on a statistical model and reference population. The third option was to present the statistical model coefficients directly (possibly converted to adjusted hazard ratios).

Dr. Snyder asked if SRTR should consider adjusting for things that have not been identified by committees as being important for organ allocation (ie, things that are not expressly included in the allocation policy). Dr. Hunsicker said it was important to acknowledge the possibility of missing something important by not including a covariate—and, to know the intent behind doing the analysis, and only look for specific outcomes of interest. Dr. Lyden supported the idea of having a list of committee goals and policies being pursued, and using that as a guide for which metrics should be considered for adjustment to best address whether the policy being explored is equitable relative across those subpopulations.

SRTR will continue to explore options to present adjusted rates to the committee as part of the simulation reporting.

Adjusting for race in patient-facing prediction tools

Dr. Lyden said that race was recently required to be removed from estimation of glomerular filtration rate (eGFR) as listing criteria for kidney transplant. Inspired in part by the discussions over the past few years about eGFR, Dr. Miller had looked at implications of removing race from the kidney donor profile index. Dr. Lyden said another part of this discussion was whether to adjust for race in OPO performance metrics discussed earlier in the meeting.

Dr. Lyden's question to the subcommittee was whether to adjust for race and ethnicity in the patient-facing prediction tools that SRTR is producing. Several of these are on the SRTR website, mostly calculators that predict outcomes on the waiting list or outcomes posttransplant. The number of patient-facing prediction tools on the website will grow under Task 5. The primary purpose of these tools is personalized information. Dr. Lyden said tools like the kidney calculator could also be used for shared decision-making between transplant candidates and their clinicians, as well as which centers to go to, and to view patient outcomes at different centers.

In deciding whether to adjust for race, the literature shows there are different ways go about making this decision, especially in the context of a prediction problem. To address if predictions are improved by adjusting for race, ethnicity, and other attributes, calibration and discrimination can be looked at. Whether predictions are improved by those metrics in each subgroup of race and ethnicity can also be looked at. A question to consider is if race is added to a model and it only changes absolute predictions by a few percentage points, is it worth including? Lastly algorithmic fairness measures can be used. This methodology assesses whether error rates (eg, false positives and false negatives) of the model predictions are equal across groups.

Dr. Lyden asked the subcommittee for opinions about whether to adjust for race in patient-facing prediction tools and more generally about how SRTR should develop overarching principles to guide our use of race in various modeling contexts. Dr. Li suggested separating the performance evaluation of programs or OPOs from personalized predictions created for patients. Dr. Hunsicker agreed with the idea of separating these two modeling paradigms. Dr. Logan asked for clarification as to how center is included in the prediction models, and Dr. Lyden clarified that center is included as a random effect in a mixed-modeling approach. The committee was supportive of continuing to use race as a prediction element in these patient-facing tools if it improved predictions but noted that SRTR should ask the patients how they view this issue. This issue will be brought before the SRTR's Patient and Family Affairs Subcommittee this coming Thursday, August 24, 2023.

Closing business

With no other business being heard, the meeting concluded. The next meeting date is to be determined.