# SRC-AMS Meeting Minutes

## Analytical Methods Subcommittee Teleconference

December 12, 2023, 12:30 PM – 3:00 PM CST

**Voting Members:**
David Vock, PhD (Co-chair)
Shu-Xia Li, PhD
Brent Logan, PhD
Erika Helgeson, PhD
Megan Neely, PhD
**Not in Attendance:**
Andrew Schaefer, PhD
William (Bill) Irish, PhD
Katherine Panageas, PhD

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

**HRSA:**
Shannon Dunne, JD
**Not in Attendance:**
Adriana Martinez, MS

**SRTR Staff:**
Larry Hunsicker, MD
Ajay Israni, MD, MS
Grace Lyden, PhD
Jon Miller, PhD
Josh Pyke, PhD
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 and nominating process

With the AMS operating on 3-year terms, Dr. Snyder thanked Dr. Shu-Xia Li and Dr. Katherine Panageas for their service, as they are finishing their terms on December 31, 2023. He said the SRTR Review Committee (SRC)–approved nominating process was implemented in September 2023, with an open call for nominations starting in early September and closing October 6, 2023. Drs. Snyder and Vock reviewed the applicants and gave their suggestions to the SRC meeting on October 27, 2023, when final recommendations were made.

There will be three incoming AMS members on January 1, 2024. The first is Dr. Joel Adler, an Assistant Professor of Surgery at the University of Austin, Texas, who has done a lot of work in the health services research domain and transplantation. The second is Dr. Syed Ali Husain, an Assistant Professor of Medicine at Columbia University Medical Center, who has done analytic work with SRTR data. The third is Dr. William F. Parker, an Assistant Professor of Medicine and Public Health Sciences at the University of Chicago, who is also the Assistant Director for the MacLean Center for Clinical Medical Ethics.

**CMS's organ procurement organization performance metrics**

Dr. Snyder gave a brief overview of organ procurement organizations (OPOs). There are 56 OPOs in the United States. Each OPO covers varying degrees of geography—some serve parts of states, whole states, or multistate regions. In 2020, the Centers for Medicare & Medicaid Services (CMS) published new performance metrics, the donation rate and transplant rate, that OPOs will be evaluated on during each year.

Both metrics have the same denominator, derived from death certificate data made available by the Centers for Disease Control and Prevention (CDC), known as the cause, age, and location consistent (CALC) death count. The numerator for the donation rate is the count of donors that had at least one organ transplanted or the pancreas sent for research. The numerator for the transplant rate is the number of organs transplanted; a pancreas sent for research counts as an organ transplanted for the transplant rate. The donation rate is not risk adjusted, whereas the transplant rate is adjusted for the age of the decedent.

CMS uses the donation rate and transplant rate to stratify OPOs into three tiers based on whether the OPO is statistically significantly below the prior year's 75th percentile (Tier 2) or the prior year's median (Tier 3) performance on at least one of the metrics. In 2026, Tier 1 OPOs will be recertified for an additional 4-year period; Tier 2 OPOs will be allowed to recompete for their contract, but other OPOs would be allowed to compete for the service area; and Tier 3 OPOs will be decertified.

Dr. Snyder said the following presentation by Dr. Jon Miller will be about a report SRTR is producing to provide OPOs with details of their performance within various subgroups and in the most recent years in which the CMS metrics have yet to be reported due to a lag in data availability from the CDC (2022 and 2023). For those most recent years, Dr. Miller has created a model to predict the denominator (CALC). This prediction will be shared with the OPOs so they can see how they are performing in the more recent years.

**Prediction of CALC deaths**

Dr. Miller noted the main issue and motivation for this project was that the CMS metrics for OPOs use CDC data that have a 2-year lag to calculate CALC deaths. He reviewed the methods that went into creating a model to predict the CALC deaths. The monthly count of CALC deaths was predicted using a mixed-effects model. Predictors included total referrals to the OPO, the number of imminent and eligible deaths referred to the OPO, and indicators for the month and year of the referral. The model included a random intercept for each OPO. The conditional $R^2$ was 0.97 with a marginal $R^2$ of 0.10. In years where the true CALC denominator was known, SRTR found that the model predictions were within about 10% of the actual CALC deaths.

SRTR thought these results were a good source of information for OPOs trying to determine their performance metrics in the most recent years. Dr. Miller reminded the subcommittee that SRTR needed clearance from the Health Resources and Services Administration (HRSA) before providing this information to the OPOs.

Dr. Megan Neely suggested SRTR provide confidence limits on the prediction to give an upper and lower bound on where OPOs may be. Dr. Vock said breaking down the data by subgroup (eg, age,

race) would help drive OPO donation improvement. Dr. Snyder noted subgroup data was part of an SRTR report on time trends that is planned to be released to OPOs on December 15, 2023. Dr. Miller spoke to the problem of being unable to predict age-adjusted transplant rates in the report given total referrals reported to the Organ Procurement and Transplantation Network (OPTN) are not disaggregated by age.

The subcommittee discussed potential pitfalls of releasing these predictions to OPOs, as this was a concern expressed by CMS. The subcommittee did not express strong concerns but suggested the inclusion of a prediction interval could add further context to interpreting the predictions. Members also suggested considering parameterizing the year as a factor rather than a linear.

**Alternative OPO flagging metric**

Dr. Grace Lyden recapped the discussion from the previous AMS meeting where SRTR expressed concerns about the current CMS tiering methodology being biased against larger OPOs. Dr. Lyden referenced the 2020 SRTR paper, "The Centers for Medicare and Medicaid Services' proposed metric for recertification of organ procurement organizations: Evaluation by the Scientific Registry of Transplant Recipients," by Snyder et al, which posited that "a performance boundary based on the 75th percentile will be biased against OPOs with more potential donors (large OPOs), and conversely biased in favor of OPOs with fewer potential donors (smaller OPO)." Dr. Lyden said this was a natural consequence of using a confidence interval and seeing whether it is significantly different from a fixed constant such as the prior year's 75th and 50th percentiles.

CMS addressed this in the CMS Final Rule published in 2020, stating confidence intervals are used to make sure threshold rates are not biased against small OPOs, which have greater variability of rates due to smaller volumes. Dr. Vock commented that using the term "biased" may be eliciting a negative reaction from CMS. Rather, he noted SRTR is pointing out that CMS is only concerned about a type I error rate, and not a type II error rate.

Dr. Lyden said SRTR evaluated the current CMS tiering algorithm under four different simulated scenarios:

1. Each OPO performs at the prior year's 75th percentile (ie, no difference across OPOs).
2. Each OPO performs at the prior year's median (ie, no difference across OPOs).
3. Each OPO has a rate that randomly varies around the prior year's 75th percentile.
4. Each OPO has a rate that randomly varies around the prior year's 50th percentile.

The results from the four scenarios were presented. The first quadrant was all OPOs performing exactly at the previous year's 75th percentile, the scenario where CMS is designed to preserve type I error. The percent of OPOs falling outside of Tier 1 remains at 5% regardless of OPO volume. The second quadrant presented results when all OPOs performed exactly at last year's median. It showed the pattern that smaller OPOs were more likely to end up in Tier 1 and not have to recompete compared with the larger OPOs. This pattern is also prevalent in the third and fourth quadrants, where true rates were allowed to vary around the prior year's 75th and 50th percentiles, respectively. Dr. Snyder pointed out the concern that when all OPOs performed at the prior year's median (ie, they should be in Tier 2), there is a 70% probability that the smallest OPO would be

classified in Tier 1 and be automatically recertified, and virtually no probability that the largest OPO would be classified in Tier 1.

Dr. Lyden said SRTR is investigating alternate tiering systems that maintain a constant type I error rate across OPO volumes while maximizing the ability to detect underperforming OPOs. SRTR developed a Bayesian observed-to-expected (O-to-E) framework to be congruent with the method used to evaluate transplant programs.

SRTR identified the optimal tiering algorithm under this Bayesian framework through a simulation study. Three scenarios were simulated:

1. No differences in performance across OPOs, ie, O-to-E = 1 for all OPOs (assessing type I error rates)
2. An OPO is underperforming by 10% relative to expected (O-to-E = 0.90).
3. An OPO is underperforming by 20% relative to expected (O-to-E = 0.80).

SRTR performed a grid search over a 2-dimensional space to define tiering rules of the general form: the probability is > *X* that the rate ratio (RR) was < *Y*. *X* and *Y* were allowed to vary and each algorithm was assigned a score to determine which algorithm preserves type I error rates across OPO volumes while maximizing power to detect underperformance.

The algorithm scoring rule was:

- The rule was penalized +0.04 points if the false flagging rate for Tier 3 was higher than 5%.
- The rule was penalized +0.03 points if the false flagging rate for Tier 2 was higher than 10%.
- The rule was penalized +0.02 points for each percentage point the Tier 2 flagging rate is less than 100% when RR = 0.9.
- The rule was penalized +0.02 penalty points for each percentage point the Tier 3 flagging rate is less than 100% when RR = 0.8.

In the distribution of the scores, a lower number was better. SRTR chose the flagging rule with the lowest score and Prob2 = Prob3, where Prob2 is the probability needed to be placed into Tier 2, and Prob3 is the probability needed to be placed into Tier 3. The optimal rule was found to be:

- Tier 3: the posterior probability that the RR is less than 0.85 is greater than 95%;
- Tier 2: the posterior probability that the RR is less than 1.0 is greater than 95%; and
- Tier 1: otherwise.

Dr. Lyden showed the results of the three simulated scenarios. The rule falsely assigned Tier 3 about 5% of the time, regardless of OPO volume, when all OPOs were truly performing the same (O-to-E = 1). Within that constraint, the rule maximizes power to detect underperformance across the range of OPO volume.

Dr. Brent Logan asked how these results compared with the CMS rules. Dr. Lyden said the simulations of the CMS system showed earlier were only under the scenario when all OPOs in fact performed the same and did not show when an OPO is performing 10% or 20% lower than expected. Under this scenario, no flagging would be ideal, because the underlying truth is that OPOs are performing the same.

Dr. Lyden said potential criticisms of these simulations may be that SRTR switched to using the typical observed-over-expected metric, and is not incentivizing that year-over-year improvement of comparing against the fixed boundary of last year's 75th percentile. SRTR's thoughts on these criticisms were that although SRTR was not making a comparison with last year's 75th percentile, a metric that compares programs with their peers *during the current evaluation period* can have benefits, such as naturally incorporating external forces like COVID-19 into expectations for performance.

Also, comparing programs with national expectations can incentivize year-over-year improvement because OPOs do tend to grow and improve, driving up expected donation rates. And any metric that compares with a fixed value (eg, previous year's median) will be subject to bias against large OPOs even when all OPOs perform the same, which is not ideal in a metric.

Subcommittee members gave additional feedback. Dr. Vock said the idea SRTR proposed may continue to pit OPOs against each other, and might not foster sharing of best practices. However, Dr. Snyder thought it would have the opposite effect, with OPOs in the same period of time fostering collaboration. One of the unique things about using the prior year's 75th percentile is that OPOs do not know what that 75th percentile target is until a year after the evaluation year due to the 2-year lag in data availability.

Dr. Logan said that when an OPO is underperforming, there is still a sizeable chance it may end up in a Tier 1 renewal. There is still a lot of heterogeneity in the Tier 1 versus Tier 2 outcomes even with an attempt to control the type I error rates. He called into question what this would mean for the implications of OPOs. Dr. Snyder said 2024 is an evaluation year, but the target boundary which is based on 2023 data is currently unknown. That 75th percentile target will not be known until the first quarter of 2025. Dr. Logan agreed it made sense to use the O-to-E metric if rates are predicted to go up.

**Flagging extrapolation beyond training data in patient-facing decision aids**

Dr. Lyden said SRTR produces model-based decision aids for patients, including the kidney waitlist calculator, the heart waitlist calculator, and the long-term outcomes app. Patients can input information to predict outcomes at different centers, and see chances of transplant and waitlist mortality. A model can be applied to any combination of covariates, including combinations not observed in the training data. Problems with this include no guarantee of correct extrapolation. It may be an impossible combination or input error. There are also problems with extrapolating for center-level predictions, such as a center not having experience listing or treating patients like them.

To address this problem, SRTR has suggested developing a method to identify when patient covariates are very different from the training data at a national and transplant-center level. Dr. Lyden said it would be beneficial to patients to have a message in these apps that states the combination of selected characteristics was not observed in the data used for modeling, the prediction quality is unknown, and predictions should be interpreted with caution.

Next, Dr. Lyden defined extrapolation as predictions outside the range of data, whereas interpolation is predictions inside the range of data. Ways to flag extrapolation include finding

implausible values of covariates, which is already implemented in the SRTR decision aids by setting allowable ranges of height, weight, and other continuous predictors. The new method would focus on flagging patients with a highly unusual *combination* of covariates. Dr. Lyden used 2D scatterplots to illustrate this point, showing that unusual combinations may occur within the allowable values of height or weight.

Dr. Lyden used the SRTR heart calculator as a motivating example. Key covariates in this model included primary diagnosis, medical urgency status and qualifying criteria, life support treatments (eg, ventricular assist device [VAD], balloon pump, extracorporeal membrane oxygenation [ECMO]), and comorbidities. There are about 43,500 possible combinations of these key covariates. However, the data set used for modeling had only just above 1,000 combinations, which supports the concern that app users might enter combinations that are unrealistic or unobserved in the training data.

Dr. Lyden summarized the proposed approach she and Dr. Nick Wood used. The first step was to define the reference distribution of the average similarity for each observation in the training data to its *k* nearest neighbors. This represents how close, on average, each point in the training data was to other points in the training data. For a new observation, calculate the average distance to *k* nearest neighbors in the training data and compare that closeness with this reference distribution. If the new observation is farther from its neighbors than the vast majority of points in the training data, the new observation is flagged as outside the applicability domain (ie, extrapolation).

Dr. Lyden showed a density plot of the average Jaccard similarity to $k = 5$ nearest neighbors, which demonstrated that 98% of training data points have an average similarity of 0.82 or higher. A new point would be out of domain if its average similarity to $k = 5$ nearest neighbors in the training data was less than 0.82 based on an *m* of 98%, because it is farther away from its neighbors than 98% of training data.

Dr. Lyden went over the tuning parameters for this approach. These included choosing a distance/similarity metric, *k* for nearest neighbors, and *m* threshold for flagging. Cross validation was used to choose these to minimize false-positive rates (flagging real data as out of domain) and maximize true-positive rates (flagging fake data as out of domain).

Dr. Lyden reviewed an application of this method with the heart calculator, using medical covariates and all two-way interactions. The Jaccard index was used as a similarity metric. Five-fold cross validation was used to select from nine combinations of *k* and *m*. Fake data were created in each held-out fold, by dividing it into fourths and modifying it to make it unrealistic using four unlikely scenarios: status 6 + ECMO (N=6 in training data), status 4 by LVAD + not on VAD (N=36), status 4 by LVAD + not on VAD + balloon pump (N=0), and on dialysis + diabetes + history of cancer (N=3).

The chosen *k* and *m* were applied to a validation dataset of N=3,436 adult heart listings from December 1, 2022, to November 30, 2023. Four hundred observations from validation data were also sampled and modified to create the four unlikely scenarios. There was a false-positive rate of 3% in the validation dataset (meaning 3% of listings were falsely flagged as out of domain) and true-positive rates of 89% to 96% for three of the four unlikely scenarios. The second unlikely scenario was flagged as out of domain only 64% of the time.

Dr. Josh Pyke said it was important to consider the pros and cons of an automated data-based approach like this to identify observations outside the domain of applicability versus trying to put in place expert rules based on what subject-matter domain experts know about what is possible. Dr. Lyden agreed it was appropriate to use expert-augmented domain of applicable methodology for the patient-facing tools.

**Closing business**

With no other business being heard, the meeting concluded. The next meeting date in January 2024 is to be determined.