Received: 7 June 2018 | Revised: 24 September 2018 | Accepted: 27 September 2018

DOI: 10.1111/ajt.15132

ORIGINAL ARTICLE

AJT

# The relationship between the C-statistic and the accuracy of program-specific evaluations

**Andrew Wey**[1] | **Nicholas Salkowski**[1] | **Bertram L. Kasiske**[1,2] | **Melissa A. Skeans**[1] | **Sally K. Gustafson**[1] | **Ajay K. Israni**[1,2,3] | **Jon J. Snyder**[1,3]

[1]Scientific Registry of Transplant Recipients, Hennepin Healthcare Research Institute, Minneapolis, Minnesota

[2]Department of Medicine, Hennepin Healthcare, University of Minnesota, Minneapolis, Minnesota

[3]Department of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota

**Correspondence**
Andrew Wey, Scientific Registry of Transplant Recipients, Hennepin Healthcare Research Institute, Minneapolis, MN.
Email: awey@cdrg.org

**Funding information**
Health Resources and Services Administration, Grant/Award Number: HHSH250201500009C

The C-statistic of the risk-adjustment model is often used to judge the accuracy of program evaluations. However, the C-statistic depends on the variability in risk for individual transplants and may be inappropriate for determining the accuracy of program evaluations. A simulation study investigated the association of the C-statistic with several metrics of program evaluation accuracy, including categorizing programs into the 5-tier system and identifying programs for regulatory review. The simulation study used data from deceased donor kidney-alone transplants for adult recipients in the program-specific reports released January 2018. A range of C-statistics was generated by changing the variability in risk for individual transplants. The C-statistic had no association with *any* metric of program evaluation accuracy. Instead, the number of expected events at a program was the most important factor. For example, Spearman's rho, which is the correlation of ranks, was −0.27 and −0.72 between the true program-specific hazard ratios and assigned tiers for programs with, respectively, <3 and >10 expected events. Presence of unadjusted risk factors did not modify the associations, although the accuracy of program evaluations was systematically lower. Therefore, the C-statistic provides no information on the accuracy of program evaluations.

**KEYWORDS**
clinical research/practice, organ transplantation in general, Scientific Registry for Transplant Recipients (SRTR), statistics

## 1 | INTRODUCTION

The Scientific Registry of Transplant Recipients (SRTR) builds risk-adjustment models for posttransplant graft and patient survival. Risk adjustment ensures that recipients with more observed comorbid conditions and/or lower-quality donors do not generate worse adjusted posttransplant evaluations.[1] Risk-adjustment models are especially important for 1-year posttransplant graft and patient survival because these outcomes are important in public reporting and regulatory review.[2-5]

**Abbreviations:** CMS, Centers for Medicare & Medicaid Services; HR, hazard ratio; MPSC, Membership and Professional Standards Committee; MSE, mean-squared error; OPTN, Organ Procurement and Transplantation Network; PSR, program-specific report; SRTR, Scientific Registry of Transplant Recipients.

The quality of risk adjustment in transplantation is commonly measured by the C-statistic.[4,6,7] The C-statistic measures the ability of the risk-adjustment models to accurately order, for example, graft failure times, and is likely popular due to its relatively intuitive interpretation. Specifically, the C-statistic is interpreted as the probability that the risk-adjustment model correctly identified the graft that failed first among 2 randomly selected recipients. For example, a C-statistic of 0.5 implies that the risk-adjustment model correctly identified the graft that failed first 50% of the time (ie, no better than a coin-flip). In contrast, a C-statistic of 1.0 implies that the risk-adjustment model always correctly identified the graft that failed first. As a consequence, in the context of 1-year posttransplant survival, a C-statistic of 1.0 implies that every graft failure prior to 1 year had a higher predicted risk of

  

failure than every graft failure after 1 year. However, from a statistical point of view, the C-statistic depends on the variability in risk for individual transplants; that is, given a correctly specified model, a high C-statistic requires more variability among recipients in, for example, the 1-year probability of graft survival than a lower C-statistic.[8]

In previous research, the C-statistic had, at best, a modest association with the accuracy of cardiovascular report cards derived from logistic regression; that is, the C-statistic provided limited information on the validity of risk adjustment.[9] The C-statistic poorly measured the quality of risk adjustment because accurate estimation of the hospital cardiovascular quality metric required accurately estimating the risk for individual patients, not ranking the individual patient risks. Equivalently, accurate estimation of the hospital cardiovascular quality metric did not depend on the underlying variability in the risk for individual patients. Instead, accurate risk adjustment relied on including important risk factors and allowing continuous risk factors to associate nonlinearly with the outcome of interest.[9] Lastly, accurate estimation of program quality depended more on the effective sample size at each program than on the C-statistic of the risk-adjustment model.[10]

Despite their dependence on the variability in risk for individual transplants, the C-statistics for the posttransplant models published by SRTR are commonly cited as a reason to distrust estimated program-specific hazard ratios (HRs) in the context of both public reporting and regulatory review.[6,7] Thus, we extended the previous research to transplantation to better inform the discussion about the importance of the C-statistic in evaluating the quality of risk adjustment. Specifically, a Monte Carlo simulation study evaluated the relationship between the C-statistic and the accuracy of estimated program-specific HRs and associated metrics (eg, accuracy of regulatory identification). The simulation study was designed to mimic, to the extent possible, SRTR's process for estimating program-specific HRs, and the study was designed with similar characteristics to the posttransplant evaluations of 1-year graft survival for deceased donor kidney-alone recipients.

## 2 | METHODS

This study used SRTR data. The SRTR data system includes data on all donors, waitlisted candidates, and transplant recipients in the United States, submitted by the members of Organ Procurement and Transplantation Network (OPTN), and has been described elsewhere.[11] The Health Resources and Services Administration, US Department of Health and Human Services, provides oversight of the activities of the OPTN and SRTR contractors.

The simulation study created synthetic data to evaluate the impact of increasing variability in recipient-level risk on the accuracy of transplant program evaluations; "recipient-level risk" was equivalent to the linear predictor in a Cox proportional hazards model. The simulation study was designed to approximate the SRTR modeling process for posttransplant graft and patient survival. Many of its characteristics were derived from the risk-adjustment model for deceased donor kidney-alone transplants in adult recipients from the January 2018 program-specific reports (PSRs). See the Supplementary Materials for a detailed description of the simulation study.

A range of C-statistics was generated by scaling the level of variability in recipient-level risk. The observed standard deviation of recipient-level risk was multiplied by a scaling factor, denoted throughout by $s$, and corresponded to the relative increase or decrease in the standard deviation of recipient-level risk observed in the January 2018 PSRs. When $s = 1$, the standard deviation of recipient-level risk was equal to the observed standard deviation from the January 2018 PSRs. The C-statistic used throughout compared the ranks of graft failure times with the linear predictors from a Cox model (ie, the C-statistic commonly used to assess risk discrimination in survival analyses).

To assess the impact of unmeasured confounders, the program-specific evaluations were also estimated in the presence of an unmeasured risk factor for each value of $s$. The unmeasured risk factor was introduced at the program level; that is, a program's transplants were considered to have systematically more or less risk due to a mechanism not identified by the risk-adjustment model. This type of unmeasured confounding likely has the largest impact on the accuracy of program evaluations because the unmeasured component of program-level risk was independent of the measured components of risk.

The accuracy of the estimated program-specific posttransplant HRs was evaluated through mean-squared error (MSE), which is the averaged squared difference between the estimated and true HRs, and is a commonly used metric for evaluating the accuracy of statistical estimators.[12-17] SRTR also categorizes posttransplant outcomes into a 5-tier system for public reporting[5,18]; therefore, we estimated the Spearman's rho, which is the correlation of ranks between the true HR and the 5-tier assignment. Similarly, to evaluate the relative accuracy of regulatory review criteria, we estimated the probability that a program identified for regulatory review had a true HR >1.25. The review criteria were from the Centers for Medicare & Medicaid Services (CMS) and OPTN's Membership and Professional Standards Committee (MPSC). Since sample size should have an important effect on program evaluation accuracy, we stratified each of the metrics by 3 categories of expected events: <3, 3 to <10, and ≥10. To account for the effect of sampling error, the program accuracy metrics were averaged over 1000 iterations of the simulation. The Supplementary Materials provide further details on calculating the metrics of accuracy.

The simulations were run in R v3.4.3[19] and used the "survival"[20] and "dplyr"[21] packages. Code for the simulation study is available at https://github.com/SRTRdevhub/C_Statistic_Github.

## 3 | RESULTS

### 3.1 | Summary measures for the simulation study

The C-statistic increased with higher variability in risk for individual transplants, even though each model was correctly specified (Table 1). The C-statistic for the scenario with variability estimated

**TABLE 1** Summary statistics of the simulation scenario without unadjusted risks

| Value of *s* | C-statistic | Expected events | Observed 1-y survival | 1-y graft survival at a percentile of risk | | | |
|---|---|---|---|---|---|---|---|
| | | | | 75th | 90th | 95th | 99th |
| 0.5 | 0.57 | 1340 | 95.0% | 94.5% | 93.7% | 93.1% | 91.9% |
| 1 | 0.64 | 1340 | 95.0% | 94.1% | 92.1% | 90.7% | 87.2% |
| 2 | 0.75 | 1342 | 95.0% | 94.1% | 89.5% | 85.3% | 73.4% |
| 4 | 0.89 | 1354 | 95.0% | 96.1% | 87.8% | 76.7% | 36.5% |
| 8 | 0.97 | 1378 | 95.0% | 99.2% | 91.2% | 68.3% | 0.5% |

The *s* value controls the level of variability in the risk of individual transplants. Each simulation scenario was specifically designed to have the same overall 1-y survival percentage.

from the current models ($s = 1$) was most similar to the value observed in the PSR model (simulated = 0.64; observed = 0.66). By design, the number of expected events and observed survival at 1 year were very similar across the range of C-statistics, although the number of expected events was slightly higher for the highest C-statistics. The true survival at 1 year for transplants at different percentiles of recipient-level risk varied significantly with the C-statistic. For example, for a C-statistic of 0.97, transplants at the 75th percentile of risk had 1-year graft survival of 99.1%, while transplants at the 99th percentile of risk had 1-year graft survival of 0.5%. While the differences in survival are dramatic and unrealistic, such C-statistics would require significantly higher variability in recipient-level risk than currently observed. Finally, the presence of unadjusted risk factors did not noticeably change the C-statistic, expected number of events, or observed 1-year survival in any of the scenarios (Table 2).

### 3.2 | Association with MSE

Within each stratum of expected events, the MSE was constant across the range of C-statistics (Figure 1). As expected, the MSE was highest for programs with <3 expected events and lowest for programs with ≥10 expected events. The pattern was similar in the presence of unmeasured risk factors. In fact, the most significant impact of unmeasured risk factors was a higher overall MSE, especially for programs with >10 expected events. Thus, the C-statistic was not associated with the MSE of program-specific HRs, even in the presence of unadjusted risks. In other words, the C-statistic provided no information on the accuracy of program-specific HRs.

### 3.3 | Association with the program assignment in the 5-tier system

Spearman's rho between tier assignment and true program-specific HRs was independent of the C-statistic (Figure 2). Similar to the MSE, Spearman's rho was strongest for large programs with no unadjusted risks (approximately −0.72), and weakest for small programs in the presence of unadjusted risks (approximately −0.27). In other words, higher-tier programs were more likely to have smaller HRs than lower-tier programs when the programs were large than when they were small. Additionally, the presence of unadjusted risks attenuated the association more for programs with a larger expected number of events (eg, Spearman's rho decreased to approximately −0.62 for large programs) but was relatively unchanged for small programs. Thus, the C-statistic provided no information on the accuracy of risk-adjustment models for assigning programs within the 5-tier system. In contrast, the accuracy of the 5-tier system increased with number of expected events at a transplant program.

### 3.4 | Association with CMS and MPSC flagging

The probability that programs flagged by CMS had true HRs >1.25 was independent of the C-statistic (Figure 3). The probabilities were approximately 68% and 95% for programs with <3 and >10 expected events, respectively. For programs flagged by the MPSC, the probability that the true HR was >1.25 surprisingly decreased for the highest C-statistic. The probabilities were approximately 52% and 80% for programs with <3 and >10 expected events, respectively. In the presence of unmeasured confounding, the probabilities decreased for both the CMS and MPSC flags, and the decrease was largest for programs with >10 expected events and smallest for programs with <3 expected events. Importantly, the probabilities were lower for the MPSC flag than for the CMS flag because the MPSC criteria are uniformly less stringent than the CMS criteria.[2]

## 4 | DISCUSSION

Correctly specified models can have low or high C-statistics. In the simulation study, the C-statistic for the same correctly specified model ranged from 0.57 for the scenario with the lowest variability in risk for individual transplants to 0.97 for the scenario with the highest variability (Table 2). This illustrates that the C-statistic cannot distinguish between a correctly specified model and the underlying variability in risk for individual transplants. Thus, the C-statistic cannot identify a correctly specified model.

Despite this limitation, the C-statistic has potential utility in comparing the performance of different models for the same data set, although such comparisons must account for the issue of overfitting with, for example, cross-validation.[12] However, even in these

**TABLE 2** Comparison of the operating characteristics of the scenarios with and without unadjusted risks

| | C-statistic | | Expected events | | Observed 1-y survival | |
|---|---|---|---|---|---|---|
| Value of *s* | Without unadjusted risks | With unadjusted risks | Without unadjusted risks | With unadjusted risks | Without unadjusted risks | With unadjusted risks |
| 0.5 | 0.57 | 0.57 | 1340 | 1340 | 95.0% | 95.0% |
| 1 | 0.64 | 0.63 | 1340 | 1340 | 95.0% | 95.0% |
| 2 | 0.75 | 0.75 | 1342 | 1343 | 95.0% | 95.0% |
| 4 | 0.89 | 0.89 | 1354 | 1353 | 95.0% | 95.0% |
| 8 | 0.97 | 0.97 | 1378 | 1378 | 95.0% | 95.0% |

The unadjusted risks had a variance equal to half the variance of the program-specific hazard ratios. The *s* value controls the level of variability in the risk for individual transplants. Each simulation scenario was specifically designed to have the same overall 1-year survival percentage.

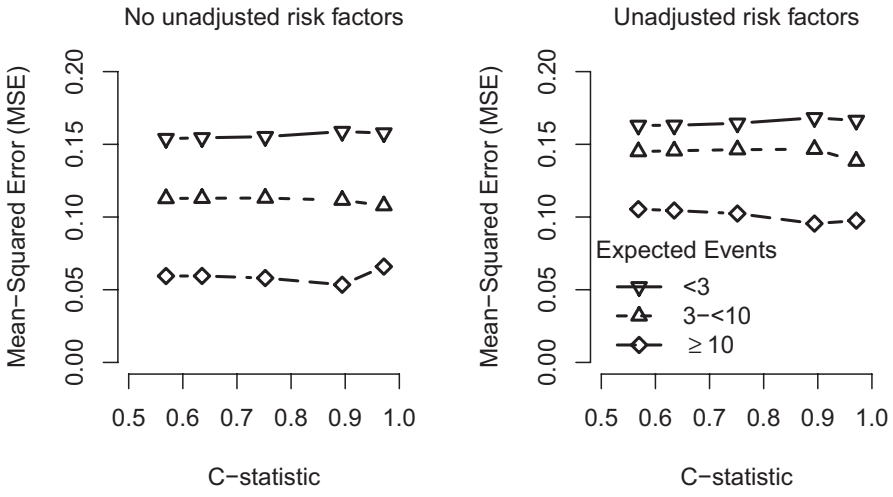


**FIGURE 1** The mean-squared error (MSE) across a range of C-statistics without (left-hand panel) and with (right-hand panel) unadjusted risk factors. Lower MSE values correspond to more accurate estimation of program-specific hazard ratios than higher MSE values

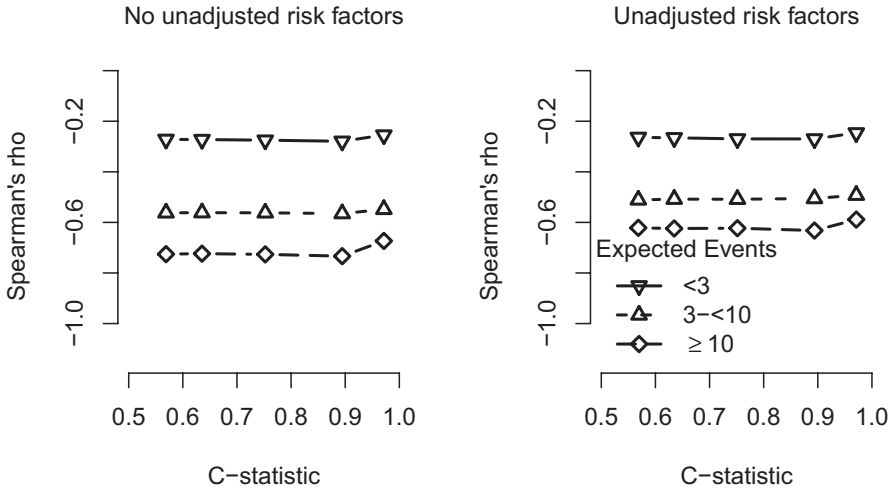


**FIGURE 2** The Spearman correlation between the tier assignment of a program and the true hazard ratio without (left-hand panel) and with (right-hand panel) unadjusted risk factors. Since higher tiers indicate better evaluations, a higher negative correlation corresponds to a stronger association between the 5-tier assignment and the true hazard ratio

situations, the C-statistic suffers from important limitations. First, the C-statistic cannot identify miscalibrated models because ranked predictions ignore the magnitude of the difference between observed and expected outcomes. Second, the traditional C-statistic used for posttransplant survival models is not guaranteed to identify the "best" model for estimating the risk of, for example, 1-year graft survival.[22] In contrast, measures of predicted error do not suffer from these limitations. For example, the Brier Score is a measure of squared error at, for example, 1 year posttransplant, and will identify both miscalibrated models and the "best" model for predicting graft survival at 1 year.

Importantly, there exists *no* measure of risk discrimination or predicted error that can identify a correctly specified model, because they all depend on unknown characteristics of the data. For example,

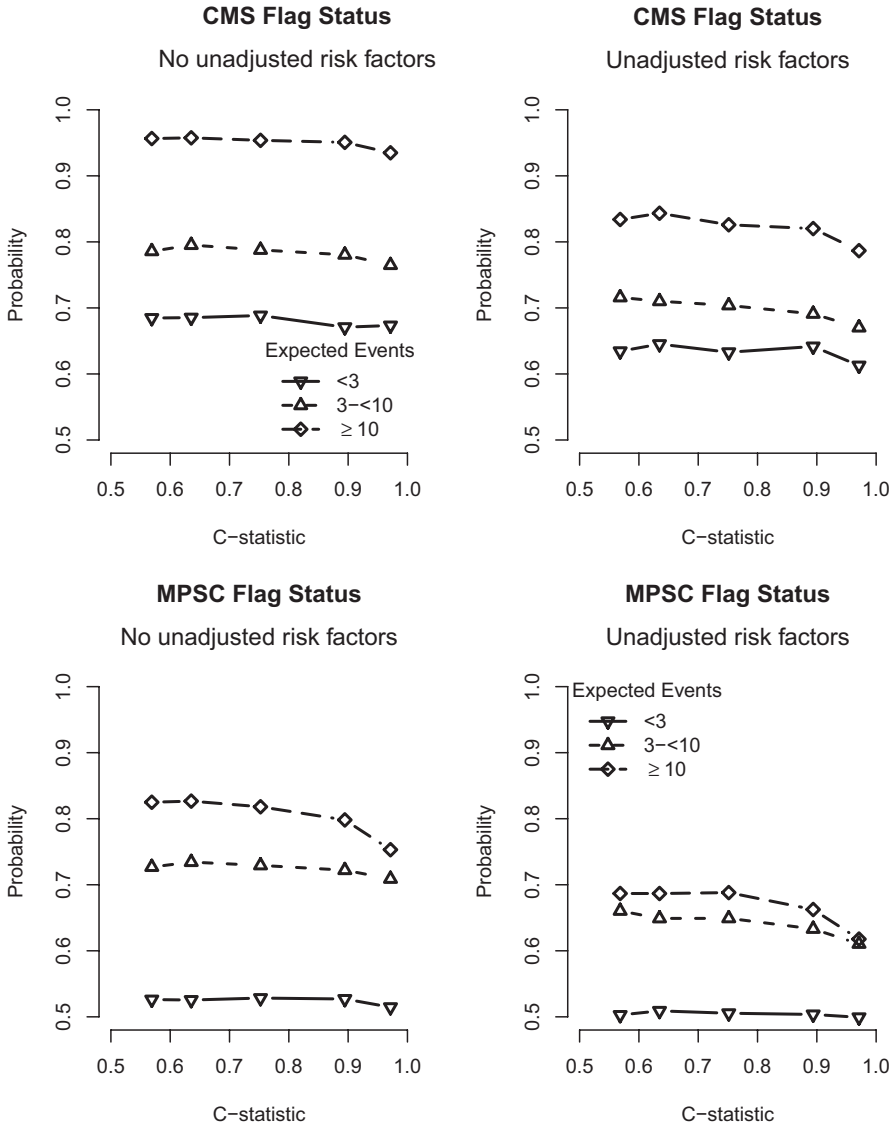


**FIGURE 3** The probability that a program flagged by the Centers for Medicare & Medicaid Services (CMS) (top panels) or the Membership and Professional Standards Committee (MPSC) (bottom panels) had a true hazard ratio >1.25 without (left-hand panels) and with (right-hand panels) unadjusted risk factors

the C-statistic depends on the variability in recipient-level risk, while measures of squared error such as the Brier Score depend on residual variability. While measures of risk discrimination and predicted error can compare the performance of 2 different models, neither can provide information on the performance of an individual model. Furthermore, such measures cannot be reliably compared across different data sets because the unknown characteristics of the data likely change. For example, a comparison of the C-statistics between posttransplant survival models and cardiovascular mortality models may only identify differences in the variability of recipient-level risk, not differences in the performance of the models.

The C-statistic was not associated with *any* evaluated metric of program evaluation accuracy. The MSE of the program HR (Figure 1), the correlation between the assigned tier and the true program HR (Figure 2), and the probability that a flagged program had a true HR >1.25 (Figure 3) did not meaningfully change depending on the C-statistic. These results are likely surprising to the broader transplant community because C-statistics are generally expected to indicate the quality of risk adjustment.[6,7,18,23] However, these performance metrics of program evaluations depend on whether the model accurately predicts individual risks, not on the variability in individual risks. Since a correctly specified model can accurately predict individual risks when variability is low, the C-statistic by itself provides no information on the accuracy of estimated program-specific HRs. Instead, accurate estimation of program-specific HRs largely depends on the sample size of the transplant program (ie, number of expected events).[10]

The simulation study also demonstrated that the C-statistic does not detect unadjusted risks. The C-statistics for simulations with and without unadjusted risk factors were not meaningfully different (Table 2). Models with low C-statistics may or may not have unadjusted risks. Models with high C-statistics may or may not have unadjusted risks. In addition, performance metrics have no meaningful relationship with the C-statistic in the presence of unadjusted risks (Figures 1, 2, and 3). Thus, even in the presence of unadjusted risks, C-statistics were not associated with *any* evaluated metric of program evaluation accuracy.

Importantly, there exists *no* measure of risk discrimination or predicted error that provides information on the accuracy of estimated

program-specific HRs derived from a single risk-adjustment model. This is fundamentally related to the fact that the measures depend on unknown characteristics of the data. Thus, correctly specified models can have low C-statistics due to low variability in recipient-level risk, or high Brier Scores due to high residual variability. However, in either situation, program-specific HRs can be accurately estimated because the model may be correctly specified.

Rather than rely on heuristic appeals to "low" C-statistics, unmeasured risk factors should be identified through a critical review of the literature with an understanding that such risk factors may be correlated with currently collected risk factors. Additionally, unmeasured risk factors may require differential distribution across programs to negatively affect the accuracy of program-specific HRs. Regardless, important unmeasured risk factors, especially factors independent of current data collection, should be brought to the attention of OPTN's Data Advisory Committee or organ-specific committees. Collection of such factors would likely improve the transplant community's faith in risk adjustment and may improve the estimation of program evaluations, especially if the factors are differentially distributed across programs.

All evaluated metrics of program evaluation accuracy were worse in the presence of unadjusted risks. The MSE increased, the tier correlation with the true HR weakened, and the probability that a flagged program had an HR >1.25 decreased. So, the simulation study does not justify refusing opportunities to build better models. Instead, it demonstrates the importance of assessing the quality of risk adjustment through a critical evaluation of the underlying statistical methodology. For example, SRTR currently builds risk-adjustment models by considering a wide range of potential risk factors and uses flexible linear splines to estimate the effect of continuous risk factors.[4] While a wide range of risk factors is considered, the modeling approach relies on the proportional hazards assumption, and accounting for nonproportional hazards may improve the risk adjustment (eg, the effect of bilateral versus single lung transplant is known to have nonproportional hazards).[24] Alternatively, better integration of interactions could also improve risk adjustment.[25] In other words, a better understanding of the effect of violated assumptions on the performance of current models would help develop better risk adjustment.

Despite following the SRTR process for estimating posttransplant program-specific HRs, this simulation study has limitations. First, survival times were simulated from a proportional hazards model. Thus, the risk-adjustment model in the simulation was correctly specified, which is unlikely to happen in practice. Second, we assumed a normal distribution of recipient-level risk. Different distributions may change the C-statistic, assuming the same level of variability in recipient-level risk, but seems unlikely to affect the association between the C-statistic and the accuracy of program-specific HRs.

We illustrated that the C-statistic of risk-adjustment models provides no information on accuracy of program-specific HRs, categorization of programs into the 5-tier system,[5] identification of programs for regulatory review, or presence of unadjusted risk factors. Instead, a program's volume (ie, the number of expected events) was the most important determinant of the accuracy of program-specific HRs.

## ACKNOWLEDGMENTS

This work was conducted under the auspices of the Minneapolis Medical Research Foundation, contractor for the Scientific Registry of Transplant Recipients, as a deliverable under contract number HHSH250201500009C (US Department of Health and Human Services, Health Resources and Services Administration, Healthcare Systems Bureau, Division of Transplantation). As a US Government-sponsored work, there are no restrictions on its use. The views expressed herein are those of the authors and not necessarily those of the US Government. AKI was partially supported by R01 HS 24527. The authors thank SRTR colleague Nan Booth, MSW, MPH, ELS, for manuscript editing.

## DISCLOSURE

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

## REFERENCES

1. Snyder JJ, Salkowski N, Wey A, et al. Effects of high-risk kidneys on Scientific Registry of Transplant Recipients program quality reports. *Am J Transplant*. 2016;16(9):2646-2653.
2. Kasiske BL, Salkowski N, Wey A, et al. Potential implications of recent and proposed changes in the regulatory oversight of solid organ transplantation in the United States. *Am J Transplant*. 2016;16(12):3371-3377.
3. Salkowski N, Snyder JJ, Zaun DA, et al. Bayesian methods for assessing transplant program performance. *Am J Transplant*. 2014;14(6):1271-1276.
4. Snyder JJ, Salkowski N, Kim SJ, et al. Developing statistical models to assess transplant outcomes using national registries: the process in the United States. *Transplantation*. 2016;100(2):288-294.
5. Wey A, Salkowski N, Kasiske BL, et al. A five-tier system for improving the categorization of transplant program performance. *Health Serv Res*. 2018;53(3):1979-1991.
6. Gupta A, Ho B, Ladner DP, Kang J, Skaro A, Kaplan B. Program-specific reports: a guide to the debate. *Transplantation*. 2015;99(6):1109-1112.
7. Jay C, Schold JD. Measuring transplant center performance: the goals are not controversial but the methods and consequences can be. *Curr Transplant Rep*. 2017;4(1):52-58.
8. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol*. 2012;12:82.
9. Austin PC, Reeves MJ. The relationship between the c-statistic of a risk-adjustment model and the accuracy of hospital report cards: a Monte Carlo study. *Med Care*. 2013;51(3):275-284.
10. Austin PC, Reeves MJ. Effect of provider volume on the accuracy of hospital report cards: a Monte Carlo study. *Circ Cardiovasc Qual Outcomes*. 2014;7(2):299-305.
11. Leppke S, Leighton T, Zaun D, et al. Scientific Registry of Transplant Recipients: collecting, analyzing, and reporting data on transplantation in the United States. *Transplant Rev*. 2013;27(2):50-56.

12. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY: Springer-Verlag; 2009.
13. Wey A, Wang L, Rudser K. Censored quantile regression with recursive partitioning-based weights. *Biostatistics*. 2014;15(1):170-181.
14. Peng L, Huang Y. Survival analysis with quantile regression models. *J Am Stat Assoc*. 2008;103(482):637-649.
15. Wang HJ, Wang L. Locally weighted censored quantile regression. *J Am Stat Assoc*. 2009;104(487):1117-1128.
16. Wey A, Vock DM, Connett J, et al. Estimating restricted mean treatment effects with stacked survival models. *Stat Med*. 2016;35(19):3319-3332.
17. Wey A, Connett J, Rudser K. Combining parametric, semiparametric, and non-parametric survival models with stacked survival models. *Biostatistics*. 2015;16(3):537-549.
18. Schold JD, Andreoni KA, Chandraker AK, et al. Expanding clarity or confusion? Volatility of the 5-tier ratings assessing quality of transplant centers in the United States. *Am J Transplant*. 2018;18:1494-1501.
19. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018. https://www.R-project.org/. Accessed June 4, 2018
20. Therneau TM. A Package for Survival Analysis in S. Version 2.38. 2015. https://CRAN.R-project.org/package=survival. Accessed June 4, 2018.
21. Wickham H, Francois R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. R package version 0.7.4. 2017. https://CRAN.R-project.org/package=dplyr. Accessed June 4, 2018.
22. Blanche P, Kattan MW, Gerds TA. The C-index is not proper for the evaluation of t-year predicted risks. *Biostatistics*. 2018. https://doi.org/10.1093/biostatistics/kxy006.
23. Axelrod DA, Friedewald JJ. Utilizing high-risk kidneys - risks, benefits, and unintended consequences? *Am J Transplant*. 2016;16(9):2514-2515.
24. Thabut G, Christie JD, Kremers WK, et al. Survival differences following lung transplantation among US transplant centers. *J Am Med Assoc*. 2010;304(1):53-60.
25. Haugen CE, Thomas AG, Garonzik-Wang J, et al. Minimizing risk associated with older liver donors by matching to preferred recipients: a national registry and validation study. *Transplantation*. 2018;102:1514-1519.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Wey A, Salkowski N, Kasiske BL, et al. The relationship between the C-statistic and the accuracy of program-specific evaluations. *Am J Transplant*. 2019;19:407–413. https://doi.org/10.1111/ajt.15132