CIBMTR®
CENTER FOR INTERNATIONAL BLOOD
& MARROW TRANSPLANT RESEARCH

Summary and Recommendations of the
2018 Center Outcomes Forum
Held on September 28, 2018

Executive Summary

The 2018 Center Outcomes Forum was held on September 28, 2018. The CIBMTR® (Center for International Blood and Marrow Transplant Research) invited representatives of the hematopoietic cell transplantation (HCT) community, the American Society for Blood and Marrow Transplantation (ASBMT) Quality Outcomes Committee, Foundation for the Accreditation of Cellular Therapy (FACT), National Marrow Donor Program (NMDP), governmental funding agencies, patients, private payers, and statisticians to participate in discussion around three key topics involving center-specific outcomes reporting:

1. Recommendations from a Pediatric Non-malignant Disease Risk Adjustment Workgroup for new variables to incorporate in the analysis
2. Recommendations from a Statistical Methodology Workgroup regarding statistical modeling
3. Managing the Consequences: How to improve collaboration to achieve quality improvement

The main discussion and recommendations for each are briefly summarized in the following pages. The final recommendations by topic include:

General Recommendations

- CIBMTR should continue to collect the patient and disease variables included in the 2018 Center-Specific Survival Analysis and evaluate their significance over the next few years.
- CIBMTR should evaluate additional quality improvement tools it can develop on behalf of centers, using input from ASBMT Quality Outcomes Committee and HCT center users.
- CIBMTR should evaluate centers' performance under the current (2018) and former risk adjustment schemes to further understand the impact of the new model on centers' performance.
- Because there are frequent misunderstandings about the Center-Specific Survival Analysis, CIBMTR should consider publishing and maintaining an FAQ page on its website.
- CIBMTR should re-engage the ASBMT Quality Outcomes Committee, representing the HCT community, to define specific and meaningful patient risk cohorts which can be used by centers to inform subgroup analyses for use in quality improvement or corrective action plans.
- The ASBMT Quality Outcomes Committee, FACT, transplant physician researchers and other stakeholder groups should define and propose research studies about topics relevant to center-specific survival analysis and public reporting to inform CIBMTR's research portfolio.

CIBMTR® is a research collaboration between
the National Marrow Donor Program® (NMDP)/
Be The Match® and the Medical College of Wisconsin

cibmtr.org

Milwaukee Campus
Medical College of Wisconsin
9200 W Wisconsin Ave, Suite C5500
Milwaukee, WI 53226 USA
(414) 805-0700

Minneapolis Campus
National Marrow Donor Program/
Be The Match
500 N 5th St
Minneapolis, MN 55401 USA
(763) 406-5800

Recommendations from the Pediatric Non-malignant Disease Risk Adjustment Workgroup for new variables to incorporate into the analysis

- CIBMTR should add the proposed data elements (Appendix C, with revisions) to the upcoming revised version of the pre-TED forms to improve risk adjustment for pediatric indications.
- The pediatric physician experts should work with the ASBMT to develop and publish a manuscript outlining appropriate pre-HCT evaluation of patients with non-malignant diseases.
- Relevant training materials for data professionals about the recommended changes should be developed, including updated information about use of the HCT-comorbidity index (HCT-CI).
- Where possible, these data elements should be tested to validate their impact on pediatric risk adjustment in the Center-Specific Survival Analysis.

Recommendations from the Statistical Methodology Workgroup regarding statistical modeling

- CIBMTR should continue to use the three indices of variability – Brier Score, R^2 score, and C-statistic to determine impact of significant variables on the multivariate modeling process.
- CIBMTR should consider using the indices of variability to test whether “center effects” significantly impact the model.
- CIBMTR should evaluate whether using more categories to further discriminate high significance variables like HCT-CI and age at HCT improves the risk adjustment model.
- CIBMTR should explore, including through research studies, whether Z-scores indicate a significant impact of “high risk” HCT recipients on center performance. The latter may influence centers’ quality improvement efforts or lead to further development of tools to support centers’ patient selection.
- Machine learning approaches should be tested to determine whether they significantly improve the risk adjustment model, and CIBMTR should consider whether to incorporate machine learning techniques into its modeling approach.

Managing the Consequences: How to improve collaboration to achieve quality improvement

- Working with ASBMT, FACT and payer representatives, a standardized process, timeline and documentation set for centers’ responses to first year performance below expected in the Center-Specific Survival Analysis should be developed.
- Addition of a short section to the ASBMT RFI that collects information about centers’ capacity, plans for expansion, innovation and research directions could improve communication with payers.
- CIBMTR should engage the ASBMT Quality Outcomes Committee, representing the HCT community, to define specific patient cohorts which can be used by centers to inform subgroup analyses for use in quality improvement or corrective action plans.
    - Provide centers with access to standardized tools through the CIBMTR Portal to perform pre-defined subgroup analyses.
- The CIBMTR Health Services and International Studies Working Committee, in collaboration with FACT and ASBMT Quality Outcomes Committee should define and propose research studies that advance our understanding of the impacts of Center-Specific Survival Analysis and public reporting on the practice of HCT.

Background

In 1986, the National Bone Marrow Donor Registry (managed by the NMDP) was established, with responsibility for maintaining an unrelated donor registry for HCT. In 1990, the Transplants Amendment Act made the reporting of center-specific outcomes for unrelated donor HCT mandatory in the United States. This activity was conducted by the NMDP from 1994 through 2007. With the Stem Cell Therapeutic and Research Act of 2005, the requirement to report HCT outcomes by transplant center was broadened to include all allogeneic (related and unrelated) HCTs in the United States. The analytic responsibility has been included in the contract for the Stem Cell Therapeutic Outcomes Database (SCTOD), currently held by the CIBMTR.

During the transition phase of the C.W. Bill Young Cell Transplantation Program (CWBYCTP), the CIBMTR, working with the NMDP, ASBMT, and the Health Resources and Services Administration (HRSA), held a meeting to review the current approach to center-specific outcomes reporting and to provide recommendations for future reports in the expanded Program. With this purpose, the CIBMTR invited representatives of the HCT community (national and international), the ASBMT Quality Outcomes Committee, governmental funding agencies, the solid organ transplant community, patients, private payers, statisticians, and experts in hospital and quality outcomes reporting to Milwaukee, Wisconsin, in September of 2008.

The objectives of the initial meeting were to review the current state of center-specific outcomes reporting in medicine and transplantation and to openly discuss strengths and limitations of current approaches with the goal of developing recommendations for HCT center-specific outcomes reports that would be:

- Scientifically valid;
- Equitable;
- Free from bias;
- Useful to the HCT community for improving quality;
- Informative for the public.

One of the recommendations of the 2008 meeting was to conduct regular reviews of the process, methodology, data collection and risk adjustment, and reporting with a broad group of stakeholders. Based on that recommendation, the Center Outcomes Forum has been held every other year since 2008 to consider the CIBMTR Center-Specific Survival Analysis. Summaries of these meetings are available at http://www.cibmtr.org/Meetings/Materials/CSOAForum.

Discussion Topics Related to Center-Specific Survival Analysis

The 2018 Center Outcomes Forum was held on September 28, 2018 and included a broad range of invited stakeholder participants (Appendix A). A summary of the group discussion and recommendations from this meeting follows.

Three workgroups were formed to present recommendations about:

- Pediatric Non-malignant Disease Risk Adjustment
- Statistical Methodology
- Managing Consequences and Improving Collaboration to Achieve Quality Improvement

Membership of the groups is shown in Appendix B, and their recommendations, as presented at the meeting, are attached in Appendix C.

Overview of 2018 Center-Specific Survival Report

Background:

An important function of the Center Outcomes Forum is to review the Center-Specific Survival Analysis and provide recommendations for improvement. It is essential that CIBMTR continue to collect relevant and updated patient, disease and transplant characteristics for use in the risk-adjustment models. While this is an ongoing process, a substantial number of data collection enhancements were made in October 2013, and those data were available to test in the 2018 report. Additionally, because this publicly available report has high impact for the HCT community, it is important to review the statistical modeling methodology to maintain accountability and transparency.

The 2018 analysis and report, which included patients who received a first allogeneic HCT between January 1, 2014, and December 31, 2016, was reviewed. More than 24,000 patients at 177 US centers met inclusion criteria. The Center-Specific Survival Analysis uses the individual center as the unit of comparison for outcomes relative to all US HCT centers. Observed one-year survival at each center is compared to the 95% confidence limit of predicted survival probability at the center based on the risk adjustment model. The multivariate risk adjustment model incorporates patient, disease and transplant characteristics to generate the predicted survival probabilities across US HCT centers using a fixed effects censored data logistic regression model to account for incomplete follow-up. The predicted survival probability at each center is the average predicted probability for all patients transplanted at the center during the time period, where the expected outcome is based on the multivariate risk adjustment model that includes all US allogeneic patients. This predicted model mimics a situation where all US HCT patients had HCT at single consolidated "generic" US center. The current multivariate model does not include a "center effect." The confidence limits surrounding the predicted survival estimate account for sample variability. More details, including descriptions of the variables themselves, can be found on the CIBMTR website.

A substantial number of additional patient and disease-related variables were available for consideration in the 2018 report risk adjustment model. The additional variables tested in the 2018 report include:

- History of mechanical ventilation
- History of invasive fungal infection
- AML transformed from Myelodysplastic (MDS) / myeloproliferative (MPN) diseases
- Therapy-related AML or MDS

- AML ELN risk group (Döhner et al. [1])
- Number of induction cycles to achieve latest complete remission (CR) before HCT for AML and ALL patients in CR
- ALL cytogenetic risk group (Moorman et al. [2])
- ALL molecular marker - BCR/ABL at any time between diagnosis and HCT
- MDS with predisposing condition
- MDS IPSS-R prognostic risk category/score at HCT (Greenberg et al. [3])
- Del 17p in CLL
- Multiple myeloma cytogenetics risk group (Palumbo et al. [4])
- Multiple myeloma International Staging System (ISS) stage at diagnosis
- Plasma cell disorder disease status at HCT
- Plasma cell leukemia
- Socioeconomic status (median household income) based on zip code of residence of recipient

Because additional data collection is burdensome for centers, CIBMTR carefully evaluated the relative contribution of these additional variables to the risk-adjustment model. The value of incorporating these risk factors was tested primarily using statistical significance of each variable in the final model. CIBMTR also assessed the relative contribution of incorporating these variables using measures of goodness of fit (including the Brier Score, R^2, and the C-statistic; discussed in the *Recommendations from Statistical Methodology Workgroup* section). New risk factors included in the 2018 analysis based on statistical significance are:

- History of mechanical ventilation
- History of invasive fungal infection
- AML transformed from MDS/MPN
- AML ELN risk group
- ALL cytogenetic risk group
- MDS with predisposing condition
- MDS IPSS-R risk score at HCT
- Plasma cell disorder disease status
- Recipient median household income based on zip code

Inclusion of these variables improved the quality of the multivariate models based on goodness of fit metrics. This suggests the collection of these data for inclusion in the Center-Specific Survival Analysis has led to measurable improvement in the risk adjustment model.

Since 2016, CIBMTR has provided analytics tools and data access for centers through the CIBMTR Portal website. All computational aspects of the Center-Specific Survival Analysis, including odds ratios for each variable and the intercept term are published with the report for transparency and for centers who wish to create additional analytic tools using their center's data.

Discussion:

There was strong endorsement for including the new variables that were statistically significant in this year's risk adjustment model, and enthusiasm for the improvements made in the modeling process.

However, there was discussion about the purpose of the report, how centers can make effective use of the information, and how the current use of the public report by payers has had negative consequences for some

HCT centers (see also “Managing the Consequences: How to improve collaboration to achieve quality improvement”). Several questions related to center outcomes reporting were elaborated.

Some attendees were concerned that small centers were at a particular disadvantage - believing a small number of patients with poor outcomes may overly contribute to poor overall performance for the center. This is accounted for using the 95% confidence limit for predicted survival as the comparison group for observed survival, where the confidence limits are greater for small sample sizes. Use of 95% confidence limits does mean there is a 2.5% probability that any given center will be determined to be performing above or below expected by chance alone, in any single year. This probability is equally distributed across centers, regardless of center size. In addition to small sample sizes, pediatric centers may care for patients with non-malignant conditions, for whom fewer disease risk adjustment variables are routinely collected. A workgroup developed recommendations for additional data collection for recipients in these disease groups which was presented at the meeting (see section *Recommendations from Pediatric Non-malignant Disease Risk Adjustment Workgroup*).

All users of the Center-Specific Survival Analysis would prefer metrics that can be used to predict future years’ performance or indicate whether a center’s observed survival is beginning to deviate from the expected survival. Centers also prefer more “real-time” performance measures to support pro-active quality improvement efforts. The current report is published nearly two years following the latest HCT episode included in the report because of inherent lags in reporting the data, the required follow-up interval, and time required to analyze and publish the report. Importantly, past center performance does not necessarily predict future performance. Centers have tools available from CIBMTR to generate predicted survivals on an individual patient basis, and examples of centers making effective use of these tools were cited. Because of the lagging nature of the formal report, many centers use their own data to inform quality improvement programs. Data from centers may be more comprehensive, offering incorporation of additional disease-specific information and analysis of additional quality metrics besides one-year survival. CIBMTR transitioned to a three-year inclusion period to provide more contemporary results in 2010. Participants suggested CIBMTR evaluate an outlier elimination method to reduce variability in center performance. This outlier elimination method would use the last five years performance, eliminate the best and worst year of the five, and produce a performance rating based on three years of data. This approach risks inclusion of less contemporary HCT, and small sample sizes at smaller centers are likely to be a limitation.

Strong interest was expressed by centers about whether patients believed to be at high risk of mortality after HCT are adequately “accounted for” in the risk adjustment model. Such high-risk patients are often treated on innovative HCT protocols designed to advance the field, and new variables were introduced to enhance the multivariate modeling this year. It is important to emphasize that each patient’s expected outcome in the model is generated based on adjustment for all risk factors included in the multivariate adjustment model, where the weight for each risk factor derives from the outcomes of all patients with the risk factor. Patients with several adverse risk factors can generally be expected to have a lower survival estimate at one year after HCT than patients with fewer adverse risk factors with the same disease and type of transplant. Any centers’ performance regarding the outcome for each patient is in comparison to that patient’s risk-adjusted predicted survival. Cumulatively, when the observed outcomes for all patients at a given center are below the lower confidence limit of the risk-adjusted predicted outcomes, the center is performing below expected. Centers providing HCT to a higher proportion of individual patients with more adverse risk factors are not inherently more likely to perform below expected. This would only be true if there are a substantial number of unmeasured factors which negatively influence outcomes that are not included in the risk adjustment model. This reinforces the importance of frequently updating data collection forms and processes with transparent and objective data elements that define essential risk factors for inclusion in the risk adjustment model.

Some payers are using results based on allogeneic HCT performance at a center to certify both autologous and allogeneic HCT for their "center of excellence (COE)" programs. There was interest in whether the center performance for allogeneic HCT can be scientifically correlated with autologous HCT outcomes. Several limitations affect CIBMTR's ability to systematically analyze autologous HCT center performance with risk adjustment, especially the relatively small number of deaths in the first year after autologous HCT, and reporting of autologous HCT is voluntary and therefore not universal among HCT centers. Risk adjustment factors, including the HCT-CI, may have different effects for autologous HCT than for allogeneic HCT. This and other similar topics represent meaningful research opportunities (see section *Managing the Consequences*).

General Recommendations:

- CIBMTR should continue to collect the patient and disease variables included in the 2018 Center-Specific Survival Analysis and evaluate their significance over the next few years.
- CIBMTR should evaluate additional quality improvement tools it can develop on behalf of centers, using input from ASBMT Quality Outcomes Committee and HCT center users.
- CIBMTR should evaluate centers' performance under the current (2018) and former risk adjustment schemes to further understand the impact of the new model on centers' performance.
- Because there are frequent misunderstandings about the Center-Specific Survival Analysis, CIBMTR should consider publishing and maintaining an FAQ page on its website.
- CIBMTR should re-engage the ASBMT Quality Outcomes Committee, representing the HCT community, to define specific and meaningful patient risk cohorts which can be used by centers to inform subgroup analyses for use in quality improvement or corrective action plans.
- The ASBMT Quality Outcomes Committee, FACT, transplant physician researchers and other stakeholder groups should define and propose research studies about topics relevant to center-specific survival analysis and public reporting to inform CIBMTR's research portfolio.

1. Recommendations from Pediatric Non-malignant Disease Risk Adjustment Workgroup for new variables to incorporate into the analysis

Background:

Although non-malignant disease indications for HCT are the minority of allogeneic HCT performed annually in the US, they represent as much as 50% of HCT performed at some pediatric centers. The distribution of these diseases are heterogenous across pediatric centers, where differential sub-specialization may exist. Collection of data for many of these diseases at the TED level has historically been limited - limiting data available for risk adjustment. To address these limitations, a workgroup was formed to recommend additional data elements to be captured on the pre-TED form. The group focused on evidence-based patient- and disease-related factors that have demonstrated impact on survival and were readily available to data professionals for reporting. Recommendations found in Appendix C were presented and discussed.

Discussion:

Recommendations for <u>all</u> pediatric patients include collection of the Glomerular Filtration Rate (GFR) before initiation of the preparative regimen, and whether the patient has known congenital heart disease (corrected or uncorrected), excluding simple atrial septal defect (ASD), ventricular septal defect (VSD) or patent ductus arteriosus (PDA) repair. The literature supporting these recommendations was reviewed, and there was consensus that these data elements affect HCT outcomes and can be collected easily by data professionals.

There has been confusion in the past about reporting co-morbid illness related to the underlying disease as part of the HCT-CI. Attendees suggested CIBMTR further clarify the appropriate use of the HCT-CI for reporting of comorbidities related to the disease for which the HCT was performed, in addition to those not related to the transplant indication.

Additional disease specific data elements for adrenal leukodystrophy, inherited erythrocyte abnormalities, and disorders of the immune system and hemophagocytic lymphohistiocytosis (HLH) were presented and discussed. There was agreement that the proposed additional data elements could be easily collected and reported by data professionals and were likely to meaningfully impact survival after HCT. The group discussed the importance of collecting liver iron concentration for patients with inherited erythrocyte abnormalities, and various testing methods. The pediatric HCT physician representatives agreed that nearly all patients would have the testing for liver iron concentration performed during routine evaluation and work-up for HCT.

Revisions were suggested to the proposed question about whether patients were colonized or infected with a viral pathogen within 60 days of HCT. The interpretation of “colonized” may be ambiguous, and alternative language – “Did the recipient have an active or recent infection with a viral pathogen within 60 days of HCT?” was proposed.

Although there was consensus that the data elements proposed for collection resulted from routine testing performed during the transplant evaluation process, a recommendation was made to collaborate with the ASBMT to develop and publish a whitepaper outlining appropriate transplant evaluation for pediatric patients and the supporting evidence. This white paper will help to inform consistent practice across pediatric HCT centers, while supporting the standard of care for insurance coverage.

In some cases, the recommended enhanced data elements are currently available at the Comprehensive Report Form (CRF)-data collection level, and if enough numbers exist these data can be used for preliminary testing of significance in a future Center-Specific Survival Analysis.

Recommendations:

- CIBMTR should add the proposed data elements (Appendix C, with revisions) to the upcoming revised version of the pre-TED forms to improve risk adjustment for pediatric indications.
- The pediatric physician experts should work with the ASBMT to develop and publish a manuscript outlining appropriate pre-HCT evaluation of patients with non-malignant diseases.
- Relevant training materials for data professionals about the recommended changes should be developed, including updated information about use of the HCT-CI.
- Where possible, these data elements should be tested to validate their impact on pediatric risk adjustment in the Center-Specific Survival Analysis.

2. Recommendations from Statistical Methodology Workgroup regarding statistical modeling

Background:

The Center Outcomes Forum is an opportunity to review the current statistical methodology and make improvements. Other than introducing new variables in the risk adjustment model, the current methodology has changed little over the last decade. However, recognizing the burden associated with data collection and reporting, questions have risen about the best way to test effectiveness of introducing new variables in the model. This is particularly relevant this year, as there is opportunity to incorporate a substantial number of new variables added to the data collection forms in 2013. Additionally, there have been questions whether CIBMTR may be missing center-based effects in its risk adjustment modeling, and whether alternative approaches to modeling, such as machine learning, may improve handling of the large number of heterogenous variables used in risk adjustment. To address these questions, a workgroup of statisticians was formed to provide recommendations (Appendix B and C).

Evaluating model performance:

The current approach to modeling uses clinical judgement as the primary criteria to determine what data to collect and test in the risk adjustment model, and which variables to include in the final models (together with levels of significance (p-values)). The workgroup recommended use of three additional measures to assess model quality. The Brier score is a measure of calibration, the Weighted C-index is a measure of discrimination, and the R^2 score is a measure of variation. All three measures use inverse probability censoring weights for the current modeling approach.

These measures were used to evaluate the degree of improvement in the 2018 Center-Specific Survival Analysis with introduction of additional patient and disease-related factors. After inclusion of the factors listed in the *Overview of 2018 Center-Specific Survival Report* section based on clinical judgement and statistical significance, all three measures were derived and showed improvement in the model. However, there was only a small change in the variability indices, suggesting there remains a substantial proportion of variability in the reason for death in allogeneic HCT recipients that is not explained by the risk adjustment model. The unexplained variation may be due to undescribed/unmeasured risk factors, or center performance. For example, the R^2 was 9.7% in 2017, and 11.2% in 2018, suggesting a small improvement and that only 11.2 % of the variability in survival is explained by the model - only a small fraction of the explainable mortality appears to be known to us based on the current model. The available information does not allow further quantification of unmeasured risk factors or center performance. Background mortality rates are not included in the center-specific modeling process, as these are assumed to be relatively consistent across US centers. There may be some variability in standard mortality rates by geographic region, however these should be accounted with the inclusion of the patient's socioeconomic status (SES) by zip code in the models.

Handling center effects:

There is heterogeneity in the type of the patients who receive HCT across centers. This "case mix" heterogeneity may represent an association between the center/provider and the risk of the patients they treat. For instance, certain high-risk patients or indications may be preferentially referred to large centers, or centers with certain characteristics or specialization. This may introduce bias in the risk adjustment model, attributable to confounding between the center effect and the patient risk effect. The risk adjustment model does not explicitly include adjustment for center effects, relying on a marginal model assumption to provide risk adjustment averaged across centers. This approach can be biased in the presence of such referral confounding, as described by recent literature in solid organ transplantation (Kalbfleisch et al. [5], Kalbfleisch et al. [6], Ash et al. [7]). Options to account for this confounding are to introduce adjustment for center effects using either a fixed effects, or random effects approach. The Statistical Methodology Workgroup thought it was essential to answer the question: "Do center effects need to be explicitly incorporated in the risk adjustment model, and if so, what is the best way?"

Strengths and weakness of the fixed effects and random effects approaches were briefly discussed. Based on workgroup recommendations, CIBMTR tested a fixed center effects risk adjustment model to compare the predicted survival with the current methodology (no center effect). Additionally, the center risk score (average risk across all patients in a center) was correlated with the center effect (estimated by a Z-score of the (observed-expected deaths)/standard error)) and with center size to further evaluate confounding. Preliminary results of the first two tests indicate minimal impact of including a fixed center effect and minimal evidence of confounding between center effect and patient risk. However, there was correlation between center risk score and center size, which may be driven by pediatric centers that are generally small and have good outcomes. This potential source of bias has a relatively small impact and may be difficult to address because there are many small centers included in the analysis.

The Statistical Methodology Workgroup will review these preliminary results, and make final recommendations about including center effects in the model and periodic testing to evaluate their impact

Alternative approaches to modeling:

The current modeling approach, using pseudo-value logistic regression modeling for one-year survival has been reliable over more than a decade. However, there are potential improvements to be considered. The data being investigated are large and heterogeneous, and handling interactions among the variables considered in the model is complicated. Traditional "manual" techniques of model building may not ascertain all relevant interactions or find the best functional form of the model to fit the data. The Statistical Methodology Workgroup considered alternative modeling approaches using machine learning techniques that may better address these challenges. Alternative modeling methods include Random Forest, Bayesian Additive Regression Trees (BART) and Stopped Cox modeling censored at one year with boosting algorithms.

Advantages and disadvantages were discussed. While alternative modeling may improve prediction accuracy of the models by incorporating data elements not considering in traditional manual models, it risks loss of transparency with stakeholders because the models are difficult to explain to clinicians. Similarly, centers may have difficulty adapting tools for use in modeling with their local data.

Discussion:

Evaluating model performance:

There were several questions about utility of the variability indices (Brier Score, R^2, C-statistic) to further understand how centers can influence their performance through quality improvement efforts. Can these variability measures be used to determine the impact of a single variable (or a small set of variables) at certain centers, or in certain disease groups? Can they be used to indicate whether there are interactions between a center and a risk factor? Unfortunately, while these evaluations are possible, small numbers of patients at individual centers will limit utility of this approach. Further, there is little indication of center effects, based on modeling done this year (see "Handling center effects").

These measures of variability provide important insight into the modeling process. While the additional patient and disease factors were statistically significant, they explain only a small additional amount of variability. This small improvement in variability is contrary to pre-conceived hypotheses about degree of importance of the added variables, especially for factors such as socioeconomic status, comorbidity, and enhanced disease risk factors for AML.

There were suggestions to use these measures of variability to examine whether removal of centers that consistently performed 'less than expected' for several years leads to improvement in the amount of explained variability in the model. This may serve to test the hypothesis that eliminating low performing centers will lead to overall improvement in HCT outcomes across the US Network, simulating the impact of eliminating centers from payer "center of excellence" networks. Unfortunately, this type of modeling does not account for whether patients would have access to HCT at other centers in such a scenario.

Handling center effects:

Interest in additional uses of the Center Risk Score and the Z-score to predict centers' performance was high. For instance, is there a difference in the outcomes of high-risk patients who have HCT at centers who care for a larger proportion of high-risk patients compared to their outcome at centers with a small proportion of high-risk patients? Can we test whether patients with higher risk of mortality are having significant effects on the center's performance? If this relationship were found to be directional, this could inform centers' quality improvement efforts. There is a generally held assumption that high-risk patients are driving center performance, but this hypothesis may be testable in a research study. Can the Z-score for center performance be used to understand centers' deviation from average over time to better predict future performance?

Questions were raised about whether the categories of certain variables used for risk adjustment provided adequate discrimination of outcome. Examples include whether the current categories of HCT-CI or age (at either extreme) sufficiently discriminate risk of death in the model.

Alternative approaches to modeling:

Stakeholder representatives agreed that concerns about transparency and local reproducibility could limit adoption of alternative modeling approaches using machine learning, and acknowledged the inherent trade-off with improvements in modeling, should they occur. There was interest in learning more about whether alternative models would explain a greater proportion of variability and improve the multivariate models.

Recommendations:

- CIBMTR should continue to use the three indices of variability – Brier Score, R^2 score, and C-statistic to determine impact of significant variables on the multivariate modeling process.
- CIBMTR should consider using the indices of variability to test whether "center effects" significantly impact the model.
- CIBMTR should evaluate whether using more categories to further discriminate high significance variables like HCT-CI and age at HCT improves the risk adjustment model.
- CIBMTR should explore, including through research studies, whether Z-scores indicate a significant impact of "high risk" HCT recipients on center performance. The latter may influence centers' quality improvement efforts or lead to further development of tools to support centers' patient selection.
- Machine learning approaches should be tested to determine whether they significantly improve the risk adjustment model, and CIBMTR should consider whether to incorporate machine learning techniques into its modeling approach.

3. Managing the Consequences: How to improve collaboration to achieve quality improvement

Background:

CIBMTR has produced the Center-Specific Survival Report to fulfill the requirements of the SCTOD as a transparent, equitable and scientifically valid performance improvement tool. Although this objective has been met, several limitations and consequences of its use are evident, particularly as the information has been used to limit centers' participation in payer network plans. Following presentations giving perspectives from relevant stakeholder groups (CIBMTR, HCT Centers, Payers, FACT and ASBMT Quality Outcomes Committee), a panel was used to facilitate discussion about actions centers and payers can take to improve collaboration and manage the consequences of the Center-Specific Survival Analysis while focusing on the common objective of quality improvement for HCT recipients.

CIBMTR perspective:

Dr. Rizzo presented the broad CIBMTR perspective about the consequences of the Center-Specific Survival Analysis. The report is considered scientifically valid, transparent and unbiased, though centers remain interested in optimizing risk adjustment. Centers and payers accept the report as high quality. Payers use the report to promote quality improvement. A substantial number of centers use the information found in the report and supplemental data provided by the CIBMTR for performance improvement efforts. The strengths of the Center-Specific Survival Analysis are its reliance on detailed clinical data, the scientific validity of the multivariate risk adjustment model, and the transparent approach.

Reliance on use of the report as a fundamental indicator of quality for HCT centers has developed, to the exclusion of other quality indicators. Some payers have used centers' performance for allogeneic HCT as a proxy indicator of centers' performance for autologous HCT when considering enrollment in COE programs. De-certification of centers from payers' COE programs has a disruptive effect on centers' referral patterns and upstream referral partners and can displace patients to programs farther from their local support networks which adversely affects patient costs, quality of life and travel for follow-up. Shifts in center referral patterns may impact local/regional program capacity at neighboring HCT centers, which may reduce timeliness of access to HCT and cause further shifting of high risk HCT candidates. Centers may increase the cost of care by performing additional testing to better document patient risk for HCT, and centers may gradually select against performing HCT in patients considered to be 'high-risk' or avoid use of innovative research protocols (which are generally undertaken in high-risk patients). Payers experience substantial disruption of their business operations and relationships.

There has been increased attention to quality improvement initiatives at HCT centers, accompanied by development of new processes and enhanced outcomes evaluation by individual HCT centers as part of the accreditation process. Recent FACT revisions require centers who perform below expected to develop corrective action plans (CAPs) to maintain FACT accreditation, and guidance from FACT for development of CAPs is increasingly sophisticated.

Center Perspective:

Dr. LeMaistre discussed the importance of using the Center-Specific Survival Analysis as part of standardization of quality improvement efforts across the Sarah Cannon Blood Cancer (SCBC) Network. A Quality Management plan has been implemented using data tools available from the CIBMTR and IT solutions developed for the network. Data driven dashboards are used to actively monitor outcomes, including unadjusted survival, non-relapse mortality, engraftment, incidence and severity of GVHD, and trigger investigation and CAPs when benchmarks are not met. Quality reporting and lessons learned are shared across all centers in the network. Many of these quality improvement systems were implemented after a center in the SCBC network had performance below expected.

FACT Perspective:

FACT accreditation standards have evolved over the last several years from process-based accreditation systems to inclusion of outcomes-based performance measurement. To achieve accreditation, centers must perform regular analysis of outcomes important to HCT and Cellular Immunotherapy and meet benchmarking requirements based on the Center-Specific Survival Analysis performed as part of the CWBYCTP. As part of this evolution, FACT's Clinical Outcomes Improvement Committee has developed standard expectations for CAPs and processes to objectively review, approve and monitor centers' CAPs when outcomes are not met. Several lessons learned during this evolution have led FACT to develop materials to educate centers about quality improvement processes and effective CAP development. FACT guidelines for CAPs include:

- Must provide specific causes of death
- Must provide quantitative data
- Must identify reasonable causes of the low one-year survival rate
- Must address the identified causes
- Must be a measurable outcome improvement
- Must provide updates at time of inspection, annual reporting, and as otherwise directed by the Committee

Dr. Gastineau provided further information about each recommendation. The Clinical Outcomes Improvement Committee issued specific guidance for centers to avoid attributing poor outcomes to perceived inadequate adjustment of high-risk patients without further attention to causes of death and development of plans to address the causes.

Following approval of a CAP by FACT, centers provide updates on their CAP through annual reports, participation in on-site inspections, and Subsequent Compliance Applications. FACT uses CAPs and updates to evaluate programs' desire to improve, implementation of the plan, and measurable improvement in one-year survival, and requests further analysis and corrective action if objectives are not met.

Payer Perspective:

Payers are generally interested in access to high quality, efficient care for the patients they cover. Although specific insurers may have slightly different focus areas, Patricia Martin, Director of Anthem's Specialty Network Development program, described Anthem's quality programs as an example of how payers make decisions about program quality. Payers place high value on transparent, standardized national data to evaluate program quality when making decisions for inclusion in covered networks, and are unlikely to be influenced by subjective data from individual centers. Like centers, payers wish to avoid disruption in their networks when centers are de-certified. Payers have relatively little independent knowledge about programs, and there are opportunities for centers to improve their communication with payers to clearly articulate their quality improvement process and analyses of outcomes, and share CAPs and timelines, as well as progress on the plans.

ASBMT Perspective:

Mark Juckett, co-chair of the ASBMT Quality Outcomes Committee, presented recommendations for collaboration. Centers should focus on creation of meaningful CAPs using the FACT process. These CAPs can serve as the foundation for collaboration between centers and payers, where robust plans that follow standardized criteria can assure quality care and provide milestones for improvement. Payers can use FACT approval of a CAP and ongoing monitoring as a prerequisite for continued access to the center for their members. Moreover, payers and centers can begin to develop processes to identify and mitigate risk factors that tend to adversely affect outcomes including late referrals for BMT, improve access to HCT centers for patients who live remote from the center, and provide programs to support patients with poor support networks related to low SES conditions. Improved payment models could be developed to provide comprehensive care coordination through the first year after HCT and beyond. Finally, centers, payers and CIBMTR can create objective criteria and reports to account for “high-risk” patients treated on clinical trials.

Discussion:

There was endorsement of the Center-Specific Survival Analysis as a high-quality report and a useful tool for evaluation of centers’ performance for allogeneic HCT. However, it represents one dimension of quality, and there was interest in understanding other measures considered to be valid by the HCT community. Aside from outcomes (which should include the ASBMT Request for Information (RFI)), other pillars of quality were suggested, including accreditations, HCT volumes, and professional team structure and credentials. FACT accreditation has evolved to include strong expectations for ongoing center performance evaluation and quality improvement processes to maintain accreditation, including development and adherence to CAPs for those centers that perform below expected. Aside from process measures, FACT expectations reinforce center’s evaluation of a range of outcomes in addition to survival, including engraftment, graft versus host disease and non-relapse mortality.

One specific aim of payer COE programs is to reinforce quality expectations for HCT centers, as well as streamline contracting. Although the probability of a center under-performing by chance alone in any single year is 2.5%, some payers have de-certified centers from their center of excellence programs based on one year of performance below expected. Aside from consequences to centers and to patients, this can cause substantial disruption to payers’ clinical operations and network relationships. All parties developed greater understanding of these impacts and expressed interest in better processes to improve quality while maintaining stability.

Some stakeholders were not familiar with the revised and strengthened FACT policies, procedures and standards in place to reinforce quality improvement efforts at centers. For those accredited centers that perform below expected in the Center-Specific Survival Analysis, there are well-defined expectations to perform self-assessment leading to meaningful action plans for improvement that must be reviewed and approved by a FACT committee. The standards for these plans were reviewed. After acknowledging the quality of the FACT CAPs, there was discussion whether these plans can form the foundation of centers’ responses to payers to address quality improvement. Payers and centers have a shared goal of high-quality patient care and the FACT CAP structure is an opportunity to inform and re-assure payers about a center’s planned quality improvement plans. A further strength of this approach is that FACT acts as a third-party in the process, providing initial review and approval of the plan, as well as providing interim monitoring and objective assessment of progress.

However, there may be additional information of benefit to payers that can supplement the FACT CAP and improve communication between payers and centers. Payers are interested in learning more about the quality tools and processes in use by centers, including processes in place to recognize quality issues in a timely manner. They are interested to learn results of root cause analysis and whether additional data available to the center provide insight into quality deficits. Payers often have very limited insight on centers’ planned growth or future clinical development plans. One constructive suggestion to improve communication between payers and centers is to add a short section to the ASBMT RFI that collects information about centers’ capacity, plans for expansion, innovation and research directions.

Recognizing that elimination of centers from excellence programs has substantial consequences for all stakeholders, there was strong support for a new collaborative process that provides meaningful reassurance regarding quality improvement. This process would use the FACT CAP as its foundation and outlines a staged approach to be taken by centers and payers. Centers that perform below expected would develop a CAP for FACT in the usual timeframe and provide early standardized communication to payers (a *Response to Concern* document) regarding their current quality processes, intended evaluation, and any interim changes in practice. Following approval of the FACT CAP, centers could share the approved CAP and timelines with payers, and provide any additional information requested by the payers. The plans and interim reports provide the opportunity for payers to develop confidence that centers are being attentive to quality improvement. With this level of assurance, payer response to a change in outcomes status for a center could include maintenance in the network while awaiting results of implementation of the FACT-approved CAP. Centers could share monitoring reports with FACT and with payers as part of the process.

Analysis of standardized patient cohorts in various risk groups defined by objective criteria may help centers evaluate their outcomes and develop action plans. These analyses may lead to specific quality improvement initiatives among patient sub-groups at centers. They may also help centers explain cohorts of patients acknowledged as having a high risk of mortality but who are being transplanted on innovative clinical studies designed to advance the field. CIBMTR and ASBMT were encouraged to define specific patient cohorts to inform such analyses.

In the course of discussion, a broad range of research questions of interest to the field were evident. Research could better define groups of patients at high risk of poor outcomes after HCT and variability in outcomes across US HCT centers. For centers that implement FACT corrective action plans, the impacts on short -term and long-term outcomes can be described. There was interest in exploring whether public reporting of center-specific survival has adversely influenced access or the types of patients undergoing HCT in the US. Analysis of enhanced datasets on selected cohorts of HCT recipients, perhaps derived from clinical trials or PRO studies, could lead to better understanding of unexplained/unmeasured sources of variability in center outcomes modeling to improve future data collection.

Recommendations:

- Working with ASBMT, FACT and payer representatives, a standardized process, timeline and documentation set for centers' responses to first-year performance below expected in the Center-Specific Survival Analysis should be developed.
- Addition of a short section to the ASBMT RFI that collects information about centers' capacity, plans for expansion, innovation and research directions could improve communication with payers.
- CIBMTR should engage the ASBMT Quality Outcomes Committee, representing the HCT community, to define specific patient cohorts which can be used by centers to inform subgroup analyses for use in quality improvement or corrective action plans.
    - Provide centers with access to standardized tools through the CIBMTR Portal to perform pre-defined subgroup analyses.
- The CIBMTR Health Services and International Studies Working Committee, in collaboration with FACT and ASBMT Quality Outcomes Committee should define and propose research studies that advance our understanding of the impacts of Center-Specific Survival Analysis and public reporting on the practice of HCT.

References

[1] Döhner H, Estey E, Grimwade D, et al., "Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel," Blood, vol. 129, no. 4, pp. 424-447, 2017.

[2] Moorman AV, Chilton L, Wilkinson J, et al., "A population-based cytogenetic study of adults with acute lymphoblastic leukemia," Blood, vol. 115, no. 2, pp. 206-214, 2010.

[3] Greenberg PL, Tuechler H, Schanz J, et al., "Revised international prognostic scoring system for myelodysplastic syndromes," Blood, vol. 120, no. 12, pp. 2454-2465, 2012.

[4] Palumbo A, Avet-Loiseau H, Olivia S, et al., "Revised international staging system for multiple myeloma: a report from International Myeloma Working Group," Journal of Clinical Oncology, vol. 33, no. 26, pp. 2863-2869, 2015.

[5] Kalbfleisch J & Wolfe RA. "On Monitoring Outcomes of Medical Providers." Statistics in Biosciences, vol. 5, pp. 286-302, 2013.

[6] Kalbfleisch J & Wolfe RA, et al., "Risk Adjustment and the Assessment of Disparities in Dialysis Mortality Outcomes." Journal of the American Society of Nephrology: JASN, vol. 26, pp. 2641-2645, 2015.

[7] Ash A, Fienberg F, et al., "Statistical Issues in Assessing Hospital Performance." https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf. 2012.

Appendix A: Attendees of 2018 Center Outcomes Forum

Full Name	Organization	Representation
Kristina Bloomquist	CIBMTR	MSP Staff
Anthony Bonagura, MD	Optum	Payer
James Bowman, MD	HRSA	Government staff
Mark Brunvand, MD	Cigna	Payer
Pintip Chitphakdithai, PhD	CIBMTR	MSP Staff
Christina Cho, MD	Memorial Sloan Kettering	HCT Center-Adult
Tonya Cox	Sarah Cannon	Center Admin
John DiPersio, MD, PhD	Washington University	HCT Center-Adult
Carol Doleysh	CIBMTR	MKE Staff
Mary Eapen, MBBS, MS	CIBMTR	CIBMTR MD
Stephanie Farnia, BA, MPH	ASBMT	ASBMT
Dennis Gastineau, MD	Mayo Clinic	ASBMT/FACT
Jeff Haertling	Consumer Advocacy Committee	Patient Advocate
Alicia Halfmann	CIBMTR	MKE Staff
Hilary Hall	Consumer Advocacy Committee	Patient Advocate
Robert Hartzman, MD	Navy	Government staff
Glenn Heller, PhD	Memorial Sloan Kettering	PhD Statistician
Mary Horowitz, MD, MS	CIBMTR	CIBMTR MD
Dianna Howard, MD	Wake Health	HCT Center-Adult
Samantha Jaglowski, MD, MPH	Ohio State University	ASBMT QOC/Adult
Mark Juckett, MD	University of Wisconsin	ASBMT QOC/Adult
Roberta King, MPH	CIBMTR	MSP Staff
Janet Kuramoto-Crawford, PhD, MHS	HRSA	Government staff
Joanne Kurtzberg, MD	Duke University	HCT Center-Peds/CBB
Michelle Kuxhausen, MS	CIBMTR	MSP Staff
Leslie Lehmann, MD	Harvard	ASBMT QOC/Peds
C. Fred LeMaistre, MD	Sarah Cannon	HCT Center-Adult
Susan Leppke, MPH	Be The Match	Be The Match Staff
Sue Logan, BS	CIBMTR	MSP Staff
Brent Logan, PhD	CIBMTR	CIBMTR PhD
Navneet Majhail, MD, MS	Cleveland Clinic	ASBMT QOC/Adult
Wendy Marinkovich, MPH	Blue Cross Blue Shield	Payer
Patricia Martin, BSN	Anthem	Payer
Richard Maziarz, MD	Oregon Health & Science University	HCT Center-Adult
Elizabeth Murphy, EdD	Be The Match	Be The Match Staff
Kristin Page, MD	Duke University	HCT Center-Peds
Ronald Potts, MD	Interlink Health	Payer

Full Name	Organization	Representation
J. Douglas Rizzo, MD, MS	CIBMTR	CIBMTR MD
Wael Saber, MD, MS	CIBMTR	CIBMTR MD
Mary Senneka	Be The Match	Be The Match Staff
Shalini Shenoy, MD	Washington University	HCT Center-Peds
Alicia Silver, MPP	Be The Match	Be The Match Staff
Steve Spellman, MBS	CIBMTR	MSP Staff
Keith Stockerl-Goldstein, MD	Washington University	HCT Center-Adult
Jesse Troy, PhD, MPH	Duke University	PhD Statistician
Julie Walz	Humana	Payer
Mei-Jie Zhang, PhD	CIBMTR	CIBMTR PhD

Appendix B: Working Group Members

Pediatric Risk Adjustment Working Group

Full Name	Organization	Representation
Stella Davies, MBBS, PhD, MD, BS (chair)	Cincinnati Children's Hospital	HCT Center-Peds
Carol Doleysh	CIBMTR	MKE Staff
Joanne Kurtzberg, MD	Duke University	HCT Center-Peds/CBB
Paul Orchard, MD	University of Minnesota	HCT Center-Peds
Kristin Page, MD	Duke University	HCT Center-Peds
J. Douglas Rizzo, MD, MS	CIBMTR	CIBMTR MD
Shalini Shenoy, MD	Washington University	HCT Center-Peds
Mark Walters, MD	Children's Hospital, Oakland	HCT Center-Peds

Statistical Methodology Working Group

Full Name	Organization	Representation
Thomas Braun, MD, PhD	University of Michigan	PhD Statistician
Pintip Chitphakdithai, PhD	CIBMTR	PhD Statistician, MSP Staff
Carol Doleysh	CIBMTR	MKE Staff
Ted Gooley, PhD	Fred Hutchinson	PhD Statistician
Glenn Heller, PhD	Memorial Sloan Kettering	PhD Statistician
Michelle Kuxhausen, MS	CIBMTR	PhD Statistician, MSP Staff
Brent Logan, PhD	CIBMTR	CIBMTR PhD
Joycelynne Palmer, PhD	City of Hope	PhD Statistician
J. Douglas Rizzo, MD, MS	CIBMTR	CIBMTR MD
Wael Saber, MD, MS	CIBMTR	CIBMTR MD
Steve Spellman, MBS	CIBMTR	MSP Staff
Jesse Troy, PhD, MPH	Duke University	PhD Statistician

Appendix C: Working Group Recommendations

Pediatric Risk Adjustment Working Group

Recommended Additions to Forms 2400/2402 to improve risk adjustment in non-malignant diseases

General – applies to all pediatric age 18 or less
Please report the measured GFR before initiation of prep regimen. xxx ml/min. [Will not collect methodology] (cutpoint TBD depending upon disease, age)
Does the patient have known complex congenital heart disease (corrected or uncorrected), excluding simple ASD, VSD or PDA repair? Yes/No

Adrenal Leukodystrophy (ALD – 543)
Please report the Loes composite score. xx (Range 0-34)
(Ref: Loes, DJ AJNR October 1994, Adrenoleukodystrophy: a scoring methodology for brain MR observations)
(Threshold score ≥ 9 for severity)

Inherited erythrocyte abnormalities (whole category), including Sickle cell and Thalassemia:
Was liver iron concentration measured? Yes/No
If yes, report method of measurement:
Liver biopsy
MR of Liver
Ferriscan R2
Other
What is the reported Liver Iron Concentration (LIC)? xx.x mg Fe/g dry weight
Is there evidence of abnormal cardiac iron deposition based on MRI of the heart at time of transplantation? Yes/No

Thalassemia only
If available, please report the Pesaro Risk Score for the patient. Options Class 1, Class 2, Class 3, NA
For patients without a Pesaro Risk Score, was hepatomegaly present with liver size greater than 2 cm below the right costal margin? Yes/No

Sickle Cell only
Was Tricuspid Regurgitant jet velocity measured by Echocardiography pre-HCT? Y/N
If yes, what was the measurement? xx m/sec

Disorders of the Immune System (includes SCID)
Is the patient colonized or infected with a viral pathogen within 60 days of HCT? Yes/No [[NOTE, we will define colonized with a polymerase chain reaction (PCR)-based definition]]
If yes, select all that apply (use ID form pick list)
Has the patient ever been infected with PCP/PJP (pneumocystis pneumonia)**?** Yes/No

SCID only
Does the patient have GVHD due to maternal cell engraftment pre HCT? Yes/No

HLH:

Is the patient colonized or infected with a viral pathogen within 60 days of HCT? Yes/No [[NOTE, we will define colonized with a PCR based definition]]

If yes, select all that apply (use ID form pick list)

Has the patient ever been infected with PCP/PJP? Yes/No

Statistical Methodology Working Group

Evaluating Model Performance

The current modeling approach uses clinical judgement to determine what data to collect and consider for testing in the risk adjustment model, along with levels of significance (p-values) for these variables as an important factor used to determine inclusion in the model. The Statistical Methodology Workgroup recommended using a combination of 3 measures to further assess model quality.

- Measure of calibration: Brier score – inverse probability censoring weights
- Measure of discrimination: Weighted C-index – inverse probability censoring weights
- Measure of variation: R^2 with inverse probability censoring weights

These measures will be tested using the 2018 Center-Specific Survival Analysis Report, over the whole model, and possibly in certain subgroups of the overall population, focusing on changes in the measures as additional variables are introduced in the model.

Handling Center Effects

Although we do not have evidence to confirm, there may be heterogeneity in "case mix" of the patients across centers. This heterogeneity may represent an association between the center/provider and the risk of the patients they treat. For instance, certain high-risk patients or indications may be preferentially referred to certain large centers, or centers with certain characteristics. This may introduce bias in the risk adjustment model, attributable to confounding between the center effect and the patient risk effect. The current risk adjustment model does not explicitly include adjustment for center effects, relying on a marginal model assumption to provide risk adjustment averaged across centers. This approach can be biased in the presence of such confounding.

Options to account for this confounding are to introduce adjustment for center effects into the model as:

- Fixed effects, (used in analysis of dialysis facilities; see Kalbfleisch and Wolfe (2013))
- Random effects, along with direct adjustment for potential confounders using center risk characteristics (assuming they are known and quantifiable), (discussed in White Paper on CMS methodology).

There are other differences between the fixed vs. random effects approach, besides how they handle confounding between center effect and patient risk. Random effects models tend to shrink estimated center effects closer together, resulting in a smaller absolute error overall; however, this is an average error rate achieved by smaller estimation error for centers with small effects, but larger error for outlier centers. Since provider analyses often focus on identification of outlying centers, fixed effects models may be more appealing due to their improved estimation of outlier center effects. The stats methodology working group recommends further exploration of the issue of confounding between center effects and patient risk to determine whether modifications to the current methodology are warranted. Specifically, they recommend the following steps to assess the impact of this potential source of confounding:

- Fit a risk adjustment model with fixed center effects and compare the predictions from such a fixed effect model with the predictions from the current risk adjustment model without fixed center effects.

- Correlate the center risk score, or average risk across patients within a center, with the center effect (estimated by a z-score of the (observed-expected)/SE), to look for evidence of potential confounding between center risk and center effect.
- Correlate the center risk score with center size, to assess another source of potential confounding between patient risk at a center and the center effect (induced by differences in center size).

The workgroup will use this information to make a final recommendation. If there is evidence of a sizable impact of confounding between center effect and patient risk, the likely recommendation will be to revise the risk adjustment model to force a fixed center effect into the model, and periodically re-evaluate. Inclusion of a fixed center effect into the risk adjustment model may require further adjustments to the model (such as switching from a pseudo-value logistic regression model for one-year survival to a stopped Cox model censored at 1 year), to improve estimation of the fixed center effects model when some of the center sizes are small. Stopped Cox Regression, which is essentially Cox Regression applied to data censored at 1 year of follow up, may have more stable convergence in the presence of small center sizes than the pseudo-value approach. Cox modeling is well-understood among stakeholders. It still relies on an assumption of proportional hazards, though this assumption only applies through the one-year time point. Small centers may still remain a challenge for generating accurate predicted confidence limits, however, even with a stopped Cox regression model.

Considerations of alternative approaches to modeling

The current modeling approach, using pseudo-value logistic regression modeling for one-year survival has been reliable over the course of more than a decade. However, there are potential improvements to be considered. The data being investigated are large and heterogeneous, and handling interactions among the substantial number of variables considered in the model is complicated. Traditional “manual” techniques of model building may not ascertain all relevant interactions or find the best functional form of the model to fit the data. The workgroup considered whether there were alternative modeling approaches using recent machine learning techniques that may better address these challenges.

The machine learning alternatives discussed for consideration include:

- Random Forest
- Bayesian Additive Regression Trees (BART)
- (Stopped) Cox modeling with boosting algorithm

Advantages: *Machine learning techniques* can process large amounts of data, and may discern patterns or inter-variable associations within the data that may not be considered based upon clinical suspicion, potentially leading to improved prediction accuracy of patient survival prognosis and better risk adjustment.

Disadvantages: Models based on machine learning are very complex and would be difficult to explain to users. The output is hard to translate into a clear understanding of risk factors and their magnitude of effect, as is currently done using OR and confidence limits. It may have the feel of a “black box” to center and payer stakeholders. This risks loss of transparency, as users would have difficulty understanding the model and the factors which influence it. As well, it would be difficult to adopt and use the model at individual centers to reproduce using local data to support decision making. Logistically, it also remains to be determined whether machine learning as a patient risk prediction model can be “plugged in” to our current approach to assessing center performance, or whether modifications would be necessary in order to use them for assessment of center performance.

The workgroup suggested performing preliminary investigation of the above machine learning techniques to determine how well they perform in our current datasets. Their value for this application would depend on the degree to which the prediction model is improved. “Goodness of fit” will be evaluated using criteria outlined in “Evaluating Model Performance.” The outcomes of the machine learning investigation will be discussed by the workgroup, and recommendations will be made based on the degree to which the model is improved. Recognizing the disadvantages that may be associated with machine learning, CIBMTR will review the benefits, and in consultation with the workgroup and the ASBMT Quality Outcomes Committee, make a final decision as to whether to incorporate machine learning techniques into the Center-Specific Survival Analysis process.

Final Recommendations from Statistical Methodology Workgroup regarding statistical modeling, updated June 1, 2019

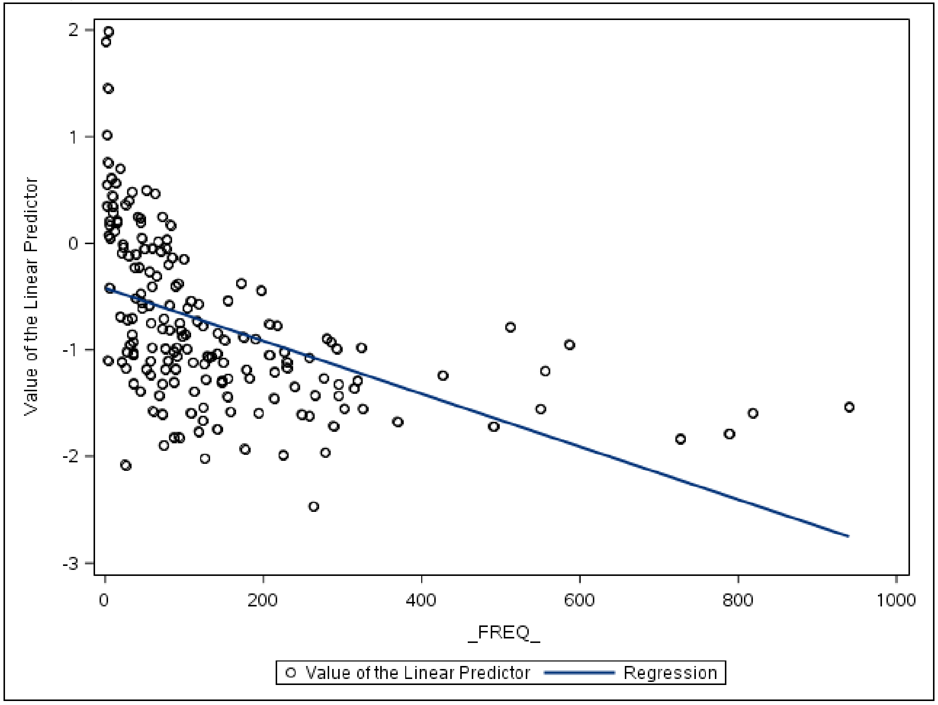
Two topics considered by the Statistical Methodology Workgroup required additional effort following the completion of the Center Outcomes Forum in September 2018. Preliminary information regarding “Handling center effects” was presented at the meeting but was planned for completion following the meeting. Additionally, there was discussion about “Alternative approaches to modeling”, specifically machine learning techniques, but analysis of these techniques was scheduled for early 2019. Final summaries of these two topics are presented below as an addendum to the Center Outcomes Forum Summary.

Handling Center Effects

The rationale for assessing bias in the risk prediction model by not including fixed center effects was discussed at the Forum and outlined earlier in this document.

The estimate of the linear risk predictor on the logit scale (Xβ term for patient characteristics only) in the risk prediction model was computed for models without a fixed center effect (as is currently done in the center outcomes analysis) and for models with a fixed center effect (as has been proposed to minimize impact of confounding between center effect and patient risk). Centers with fewer than 20 patients were collapsed together in order to be able to get the pseudo-value regression fixed effect model to converge. Estimates of the risk predictor with vs. without center effects were plotted, indicating strong agreement. The Pearson correlation between the model with and without center effects is r=0.99568, with an R^2 of 0.9914. The predicted 1-year OS probability with no center effect is highly concordant with the predicted survival probability when a fixed center effect is included. This suggests the results are virtually indistinguishable whether a fixed center effect is included.

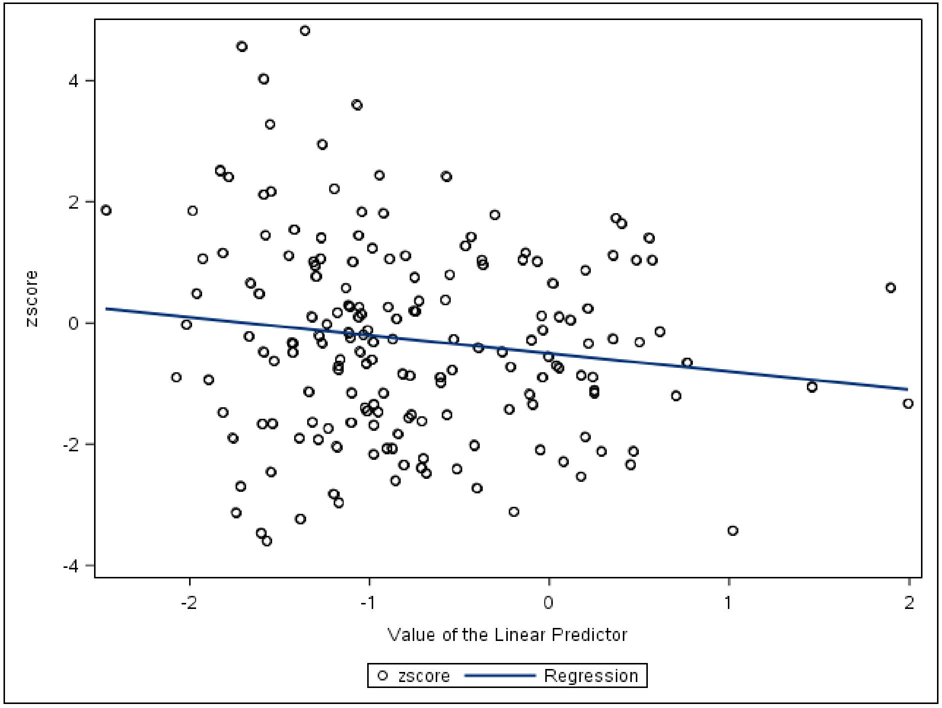
Plot of center risk score (Xβ) vs. center size



The best fitting regression line between the two is Y = (-0.42067) - 0.00248X, where X is the center size and Y is the center risk score. There is a statistically significant association between X and Y (p<0.001).

This suggests that small centers have better predicted survival. This effect may be driven by pediatric centers, which are generally small and whose patients have better survival.

Plot of z-score for center performance compared to center risk score (Xβ)



The best fitting regression line between the two is Y = (-0.50375)-0.2981 X, where X is the center risk score and Y is the center performance expressed as the Z score of (observed-expected)/SE. There is not a statistically significant association between X and Y (p=0.055).

Discussion

This helps to inform the question often asked by clinicians – "Do centers treating higher risk patients generally have worse performance in the model?" There is no statistically significant evidence that centers treating high-risk patients are more likely to be considered "under-performing" in the risk adjustment model. This indicates that our current risk adjustment is adequately accounting for patient risk. In utilizing the risk adjustment model for determining center performance, centers who do more high-risk patients are benchmarked against a different expected outcome that is matched to the risk of their patients. Therefore, they are not unfairly "penalized" for doing more high-risk transplants, because that risk is being accounted for by the model.
This graph (*Plot of z score for center performance compared to center risk score*) may also serve a future purpose as a diagnostic tool for the quality of the risk adjustment model. Systematic under- or over-representation of risk would be identifiable as a pattern in this plot.

Conclusions

- There is no evidence that addition of fixed center effects to the current risk adjustment model leads to improvement in the model.

Recommendations

- Modeling using a fixed center effect adds complexity to the modeling, and the results of the model with and without fixed effects are indistinguishable, the Statistical Methodology Workgroup recommends no change in the current methodology which does not incorporate center effects.

- CIBMTR should re-test center effects every three years. This interval is consistent with the turnover of the patient population considered in the model and will account for new variables introduced over time.

Alternative approaches to modeling

The current modeling approach, using pseudo-value logistic regression modeling for one-year survival has been reliable over the course of more than a decade. However, there are potential improvements to be considered. The data being investigated are large and heterogeneous, and handling interactions among the substantial number of variables considered in the model is complicated. Traditional "manual" techniques of model building may not ascertain all relevant interactions or find the best functional form of the model to fit the data. The workgroup considered whether there were alternative modeling approaches using recent machine learning techniques that may better address these challenges. These techniques and their advantages and disadvantages were discussed at the Center Outcomes Forum. Subsequently, CIBMTR performed analyses using data from the 2018 Center-Specific Survival Analysis to evaluate whether use of machine learning could significantly enhance the risk adjustment modeling process.

Analysis results

The 2018 Center-Specific Outcomes dataset was updated to incorporate additional follow-up and minimize loss to follow-up prior to one year (n=163 surviving patients with <9 months follow up removed, <1%) so that simpler binary outcome prediction models could be studied. A random subset of 15% of the data (n=3545) was held out for validation and assessment of prediction performance of the various methods. Various prediction model approaches were built using the training dataset of n=20443. Three approaches were considered: Random forests, Bayesian Additive Regression Trees (BART), and Gradient Boosting and compared to logistic regression using the factors from the current report (current modeling approach but refit with the training data). For random forests, both a default version as well as a version with cross-validation of the number of trees and the mtry parameter (number of variables randomly selected as candidates for splitting a node). For BART, both a default setting and a version with cross-validation of the k parameter and the number of trees were considered. For gradient boosting, tuning parameters were selected by cross-validation. Brier scores and C statistics from applying the fitted model to the independent training set are shown in the table below for the various methods. Lower Brier scores are better, as are higher C-statistics.

Method	Brier Score	C-statistic
Logistic regression (**current model**)	**0.188**	**0.6928**
Random Forests (Default)	0.1958	0.6524
Random Forests (Cross validation)	0.1915	0.6737
BART (Default)	0.1885	0.688
BART (Cross validation)	0.1889	0.6876
Gradient boosting (Default)	0.1997	0.5069
Gradient boosting (Cross validation)	0.1874	0.6948

Conclusions

- Based on the Brier Score and C-statistic, as well as the plots, the three tested machine learning algorithms do not appear to improve the prediction of the current logistic regression model. The cross-validated gradient boosting model marginally beat the logistic regression model, but the incremental benefit is quite small.
- There may be opportunity to test additional machine learning models in the future, as this is a field which is evolving quickly. This may include Xg boosting, or deep learning methods.

Recommendations

- Since there is no significant enhancement of the prediction accuracy through the use of machine learning techniques compared to the current logistic regression modeling, and machine learning is less transparent, the Statistical Methodology Workgroup recommends continuing to use the current methodology.
- CIBMTR should re-evaluate new methods of machine learning or deep learning on a regular basis as new methods emerge and effectiveness improves.