
Benjamin S. Wessler, MD, MS a,b, Andrew R. Weintraub, MD a, James E. Udelson, MD b, and David M. Kent, MD, MS b

aDivision of Cardiology and the CardioVascular Center, Tufts Medical Center, Boston, Massachusetts, USA; bPredictive Analytics and Comparative Effectiveness (PACE) Center, Tufts Medical Center, Boston, Massachusetts, USA

ABSTRACT

Background: One third of high- and prohibitive-risk TAVR patients remain severely symptomatic or die 1 year after treatment. There is interest in identifying individuals for whom this procedure is futile and should not be offered.

Methods: We performed a systematic review of the highest reported stratum of risk in TAVR clinical predictive models (CPMs). We explore whether currently available predictive models can identify patients for whom TAVR is futile, based on a quantitative futility definition and the observed and predicted outcomes for patients in the highest stratum of risk.

Results: Seventeen TAVR CPMs representing 69,191 treated patients were published from 2013 to 2018. When reported, the median number of patients in the highest stratum of risk was 569 (range 1 to 1759). Observed mortality for this risk stratum ranged from 9% at 30 days to 59% at 1 year after TAVR. Statistical confidence in these observed event rates was low. The highest predicted event rates ranged from 11.0% for in-hospital mortality to 75.1% for the composite of mortality or high symptom burden 1 year after TAVR.

Conclusion: No high-risk TAVR group in currently available TAVR CPMs had an appropriate event rate and adequate statistical power to meet a quantitative definition of futility.

ARTICLE HISTORY Received 7 November 2019; Revised 24 April 2020; Accepted 14 May 2020

KEYWORDS Clinical cardiology; interventional cardiology; noninvasive and minimally invasive cardiology

Introduction

Recently, transcatheter aortic valve replacement (TAVR) has revolutionized the care of older adults with high operative risk or those previously considered inoperable, enabling treatment of symptomatic aortic stenosis where once there was none. While most patients benefit from TAVR, one third of high- and prohibitive-risk TAVR patients remain severely symptomatic or die 1 year after treatment. Substantial work has focused on evaluating the hazard associated with various clinical factors in an effort to improve prediction, enhance shared decision-making, and identify patients who are unlikely to do well. There is an open question about whether these markers can be used to identify patients who do not benefit from TAVR. Single risk factors viewed in isolation are unlikely to effectively identify patients with extreme risk. Multivariable regression can more effectively risk-stratify patients and might be able to identify patients for whom TAVR is futile.

TAVR can be offered to patients with extremely high predicted risk who are unlikely to do well. The question of whether TAVR should be offered to these patients is in some ways harder to answer. Given the resource-constrained realities of contemporary health care and the importance of delivering care that is concordant with patient’s wishes, there is substantial interest in identifying individuals for whom this procedure is futile and should not be offered. This concept has been difficult to define because of the inherent value-laden nature of this idea and resulting lack of consensus about criteria and thresholds. A quantitative definition of futility, originally proposed by Schneiderman et al., attempted to use empirical evidence to anchor decisions about limiting treatments instead of opinion alone. This threshold was defined where “physicians [can] conclude (either through personal experience, experiences shared with colleagues or consideration of reported empiric data) that in the last 100 cases, a medical treatment has been useless, they should regard that treatment as futile.” This quantitative definition has never been explored in the case of TAVR.

We recently reported on a systematic review of CPMs for patients with valvular heart disease (VHD). This work demonstrated that there are a number of CPMs that are available to predict outcomes for TAVR patients, though these models have not been widely tested in external validations and as a result there remain substantial uncertainties about TAVR predictions. Given the inherent methodological
strengths of multivariable risk prediction, here we explore whether currently available TAVR CPMs can help identify patients for whom TAVR is futile, based on the oft-cited quantitative definition and the observed and predicted outcomes for the highest-risk patients.

Materials and methods

Systematic Review. The results of our systematic review have been previously reported. Briefly, we included data from the Tufts Predictive Analytics and Comparative Effectiveness (PACE) CPM Registry (www.pace.tuftsmedicalcenter.org/cpm), a database of predictive models for patients with cardiovascular disease. The methods for the Registry covering CPMs from January 1990 through 2015 have been previously published. The search terms used to identify CPMs are shown in the Supplemental Table. Here, we extended the PubMed search for TAVR CPMs to October 31, 2018. Citations were reviewed to confirm completeness of our review. All data fields were extracted in duplicate to ensure accuracy. Discrepancies were discussed to arrive at consensus. For inclusion in this analysis, articles had to meet the following criteria: (1) develop a predictive model containing at least two predictors as a primary aim, (2) contain a model predicting the development of a clinical outcome following TAVR, and (3) present enough information to estimate the probability for an individual patient. Articles were excluded if they did not provide enough information to predict a patient’s risk or if the described models predicted surrogate outcomes. Consistent with prior work, we excluded non-English reports, pharmacology reports, cost-effectiveness models, decision-analysis models, systematic reviews, and editorials.

Definition of Futility: We use the quantitative definition of futility proposed by Schneiderman and colleagues. The statistical basis for this determination is that given 100 successive treatment failures for patients with a certain level of risk “the clinician can be 95% confident that no more than three successes would occur in every 100 comparable trials.” Here, we explore whether a pre-procedure stratum of risk can be identified where at least 100 patients all die after being treated with TAVR. By using a statistical definition of this concept, the intent is to minimize reliance on value-laden concepts where consensus is unlikely to be achieved.

Data Extraction: For each TAVR CPM we extracted information on predicted and observed outcome rates for the highest reported stratum of risk. This risk stratum was identified either as a pre-specified risk grouping or from an assessment of the highest-risk quantile reported during assessment of model calibration. The 95% confidence interval (CI) for observed event rates in the risk stratum of interest was calculated using the exact binomial method. We compared the highest reported event rates to the quantitative definition of futility proposed by Schneiderman.

Results

Seventeen TAVR CPMs representing 69,191 treated patients were published from 2013 to 2018 (Table 1). The median number of patients used to derive these predictions was 2137 (IQR 1488). These CPMs predict mortality (N = 14) or the composite outcome of mortality or high symptom burden (N = 3). Only one CPM has the stated goal of identifying procedural futility, defined as the composite of 1-year mortality, stroke, or lack of functional-class improvement and repeat admissions occurring >1 month after the procedure.

Thirteen (76%) CPMs present risk information specific to the highest-risk stratum. These risk estimates were most frequently identified by reviewing the 10th decile of the Hosmer-Lemeshow calibration plot (N = 6). Nine CPMs (53%) did not present enough information to calculate a CI around the observed event rate. When reported, the median number of patients in the highest-risk group was 569 (range 1 to 1759). The observed mortality rates in the highest-risk group ranged from 9% (51/569; 95% CI 6.7% to 11.6%) to 100% (1/1; CI 2.5% to 100%) for mortality at 30 days. Observed mortality at 1 year for the highest-risk stratum ranged from 31% (235/763, 95% CI 27% to 34%) to 59% (19/32, 95% CI 41% to 76%). The highest predicted event rates ranged from 11.0% for in-hospital mortality to 75.1% for the composite of mortality or high symptom burden 1 year after TAVR. Statistical confidence was lowest for the highest observed event rates (Figure 1). No high-risk TAVR group had an appropriate event rate and adequate statistical power to meet a quantitative definition of futility.

Discussion

The main finding from this study is that there is no evidence that aortic stenosis patients with little opportunity for the benefit from TAVR can be identified ex ante. A review of the currently available evidence from TAVR CPMs demonstrates that these models cannot support a quantitative claim of futility for patients being considered for TAVR. Additionally, there is substantial uncertainty about outcomes for the highest-risk patients, owing to the small number of patients included in these risk strata. Taken together, these data strongly suggest that providers cannot (and should not) look to extant predictive models alone to understand whether treatment with TAVR is futile.

There are certainly some patients with symptomatic aortic stenosis who should not be treated with TAVR. Procedural denial often results from an anticipated very low likelihood of benefit to either length of life or symptom improvement, though it is unrealistic to look toward CPM-based predictions to support such a claim. A recent expert consensus document suggested a Society of Thoracic Surgeons Predictive Risk of Mortality (STS-PROM) score >15% be used as a threshold above which TAVR might be withheld based on a lack of appreciable benefit in all-cause mortality at 5 years in the PARTNER IB trial. We believe the rationale supporting this claim is flawed. First, STS-PROM discriminatory performance is substantially attenuated for patients receiving TAVR as it was not developed on these patients. Second, there were only 38 patients with STS-PROM >15% in the TAVR treated arm of PARTNER IB and 6 were still alive at 5 years. As a result, using these data alone, it is not possible to support a quantitative futility claim. Third, outcome rates from the original TAVR trials are likely not reflective of contemporary practice as devices and procedural techniques have improved over time leading to safer procedures and lower event rates. Lastly, it is clear that many patients pursue treatment with TAVR with the hope of improving symptoms – a patient-centered goal that may be achieved in the absence of mortality benefit. Ultimately, in light
of the available data, decisions for or against treatment for the highest-risk patients must take a broader view.

Multidisciplinary heart team (MDT) decisions about the appropriateness of TAVR for the highest-risk aortic stenosis patients should start with understanding procedural complexity and patient’s goals of treatment.36,37 Efforts to enhance shared decision-making and understand the likelihood of achieving patient’s goals should assume a central role.36 For the patients who are least likely to do well, a wider view of the MDT is appropriate, with the inclusion of palliative care providers before treatment to more fully explore patient’s goals, especially if TAVR is being considered primarily to improve symptoms.37 Ultimately, procedural denial may be most appropriate coming from the MDT (generally) as opposed to the operator (in isolation).11

There are limitations to this analysis. These TAVR CPMs overwhelmingly focus on mortality outcomes. As a result, improvements in symptoms are generally not accounted for. With the exception of a few CPMs,7,13 improvement in symptoms may be an acceptable outcome for many patients despite a lack of perceived mortality benefit. Another limitation to the presented data is that these CPMs are derived exclusively on patients who are treated with TAVR. In this analysis, there are no medically treated groups to compare outcome rates in the absence of TAVR though it is well known that mortality rates in untreated patients are extreme.38 Lastly, the quantitative definition of futility that is explored here may not be appropriate for TAVR decision-making, where health systems and providers may support pursuing treatment even for a very small chance of benefit (to either symptom burden or mortality).

**Conclusion**

In the case of TAVR, CPMs have assumed a central role in risk communication and have tremendous potential to enhance shared decision-making. However, currently available tools have insufficient statistical power to identify patients for whom TAVR is futile and there is substantial outcome uncertainty associated with these predictions for the highest-risk patients. More work is needed to

---

**Table 1. TAVR clinical predictive models.**

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Outcome</th>
<th>Model N</th>
<th>Events</th>
<th>Method</th>
<th>C-statistic</th>
<th>Risk Levels</th>
<th>N Highest-risk group</th>
<th>Predicted Risk (high)</th>
<th>Observed Risk (high)</th>
<th>Events (high)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kotting16</td>
<td>2013</td>
<td>In-hospital</td>
<td>11147</td>
<td>416</td>
<td>Logistic regression</td>
<td>0.81</td>
<td>Low, moderate, high</td>
<td>1038</td>
<td>0.173</td>
<td>0.17</td>
<td>176</td>
</tr>
<tr>
<td>Arnold17</td>
<td>2014</td>
<td>6-month poor</td>
<td>2137</td>
<td>704</td>
<td>Logistic regression</td>
<td>0.66</td>
<td>Low (&lt;25%), intermediate (25% to &lt;50%), high (&gt;50%)</td>
<td>-</td>
<td>0.57</td>
<td>0.61</td>
<td>NR</td>
</tr>
<tr>
<td>Arnold17</td>
<td>2014</td>
<td>1-year poor</td>
<td>2130</td>
<td>1073</td>
<td>Logistic regression</td>
<td>0.66</td>
<td>Low (&lt;25%), intermediate (25% to &lt;50%), high (50% to &lt;70%), very high (70%)</td>
<td>-</td>
<td>0.75</td>
<td>0.73</td>
<td>NR</td>
</tr>
<tr>
<td>Capodanno18</td>
<td>2014</td>
<td>30-day mortality</td>
<td>1256</td>
<td>77</td>
<td>Logistic regression</td>
<td>0.73</td>
<td>Low, medium, high</td>
<td>104</td>
<td>-</td>
<td>0.23</td>
<td>24</td>
</tr>
<tr>
<td>D’Ascenzo9</td>
<td>2014</td>
<td>30-day mortality</td>
<td>1064</td>
<td>60</td>
<td>Logistic regression</td>
<td>0.66</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D’Ascenzo9</td>
<td>2014</td>
<td>1-year mortality</td>
<td>1064</td>
<td>165</td>
<td>Logistic regression</td>
<td>0.68</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jung60</td>
<td>2014</td>
<td>30-day mortality</td>
<td>2552</td>
<td>253</td>
<td>Logistic regression</td>
<td>0.67</td>
<td>-</td>
<td>1</td>
<td>0.69</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Debonnaire21</td>
<td>2015</td>
<td>1-year mortality</td>
<td>509</td>
<td>80</td>
<td>Cox regression</td>
<td>0.715</td>
<td>Low (&lt;3 pts), high (&gt;3 pts)</td>
<td>12</td>
<td>0.60</td>
<td>0.58</td>
<td>7</td>
</tr>
<tr>
<td>Edwards22</td>
<td>2016</td>
<td>In-hospital</td>
<td>13672</td>
<td>730</td>
<td>Logistic regression</td>
<td>0.67</td>
<td>-</td>
<td>-</td>
<td>0.11</td>
<td>0.13</td>
<td>NR</td>
</tr>
<tr>
<td>Schiller16</td>
<td>2017</td>
<td>In-hospital</td>
<td>18054</td>
<td>817</td>
<td>Logistic regression</td>
<td>0.74</td>
<td>-</td>
<td>1759</td>
<td>0.16</td>
<td>0.16</td>
<td>281</td>
</tr>
<tr>
<td>Forcillo23</td>
<td>2017</td>
<td>30-day mortality</td>
<td>361</td>
<td>21</td>
<td>Logistic regression</td>
<td>0.74</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Zusman24</td>
<td>2017</td>
<td>Procedural futility +</td>
<td>435</td>
<td>66</td>
<td>Logistic regression</td>
<td>0.73</td>
<td>Low, medium, high</td>
<td>-</td>
<td>-</td>
<td>0.59</td>
<td>NR</td>
</tr>
<tr>
<td>Hermiller25</td>
<td>2016</td>
<td>30-day</td>
<td>2482</td>
<td>144</td>
<td>Cox regression</td>
<td>0.76</td>
<td>Low (Q1), moderate (Q2, Q3), high (Q4)</td>
<td>569</td>
<td>-</td>
<td>0.09</td>
<td>51</td>
</tr>
<tr>
<td>Hermiller25</td>
<td>2016</td>
<td>1-year all-cause death</td>
<td>2482</td>
<td>566</td>
<td>Cox regression</td>
<td>0.83</td>
<td>Low (Q1), moderate (Q2, Q3), high (Q4)</td>
<td>763</td>
<td>-</td>
<td>0.31</td>
<td>237</td>
</tr>
<tr>
<td>Hakkin26</td>
<td>2016</td>
<td>30-day mortality</td>
<td>1327</td>
<td>45</td>
<td>Logistic regression</td>
<td>0.63</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>NR</td>
</tr>
<tr>
<td>Lindman27</td>
<td>2015</td>
<td>1-year all-cause mortality</td>
<td>2180</td>
<td>471</td>
<td>Cox regression</td>
<td>0.65</td>
<td>Score 0–1, 2–3, 4–5, ≥ 6</td>
<td>32</td>
<td>-</td>
<td>0.59</td>
<td>19</td>
</tr>
<tr>
<td>Martin28</td>
<td>2018</td>
<td>30-day mortality</td>
<td>6339</td>
<td>326</td>
<td>Logistic regression</td>
<td>0.66</td>
<td>-</td>
<td>-</td>
<td>12.3</td>
<td>12.1</td>
<td>NR</td>
</tr>
</tbody>
</table>

*Note. Model N is total number of patients. Events represents total number of outcomes. Method represents regression method used. N Highest-risk group indicates the number of patients in the highest-risk quantile used for analysis. *defined as mortality, KCCQ-OS <45, or decrease of >10 points on KCCQ-OS from baseline. + defined as a composite of 1-year mortality, stroke, lack of functional-class improvement (by New York Heart Association class), and readmissions (>1 month after the procedure).
align the decision for or against treatment of the highest-risk individuals with patient-defined treatment goals.

**ORCID**

Benjamin S. Wessler  http://orcid.org/0000-0002-8652-1426

**Funding**

National Institutes of Health (R03 AG056447 GEMSTAR Grant and K23 AG055667) and Bellows Foundation Grant from the American College of Cardiology.

**Disclosure statement**

The authors have no conflicts of interest to report.

**References**


