RISK ADJUSTED MORTALITY FOR PTOS DATA

Douglas Wiebe PhD, Kit Delgado MD MSPH, Daniel Holena MD FACS, Brendan Carr MD MS
Department of Biostatistics and Epidemiology
Department of Emergency Medicine
Division of Traumatology and
Surgical Critical Care,
Department of Surgery
Perelman School of Medicine
University of Pennsylvania

October 25, 2014

A report prepared for the Pennsylvania Trauma Systems Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

- A risk-adjusted model was developed for the Pennsylvania Trauma Systems Foundation (PTSF) to enable comparisons of performance across trauma centers in PA.
- Performance was defined as the rate of death that was observed relative to the rate of death that was expected (O/E ratio) at each trauma center after controlling for differences in case mix and injury severity.
- The modeling methods were based upon the procedure recommended by the Trauma Quality Improvement Project (TQIP) of the American College of Surgeons.
- The analysis was conducted using Pennsylvania Trauma Outcomes Study (PTOS) data on 100,278 patients age 16 and older who were treated for blunt or penetrating trauma in one of 27 trauma centers in PA during 2011–2013.
- The final statistical model used logistic regression to predict the risk of death after adjustment for a total of 16 covariates.
- The statistical model fit the data well as indicated by conventional diagnostics including a c-statistics (an area under the receiver operating characteristic (AUC) curve) of 0.941.
- The c-statistic varied moderately among patient subgroups and was 0.932 in blunt trauma patients and 0.980 in penetrating trauma patients.
- Overall results: 2 of the trauma centers had higher than expected mortality and 5 trauma centers had lower than expected mortality.
- Subgroup results: There was variability in the number of trauma centers that had higher or lower expected mortality specifically for blunt trauma patients, multisystem trauma, head trauma, and geriatric trauma. No trauma centers had higher or lower than expected mortality for penetrating trauma.

INTRODUCTION

The goal of this project was to evaluate the performance of trauma centers in Pennsylvania (PA) and to enable comparison of performance between trauma centers. The product, provided in the form of this report, is being delivered to the Pennsylvania Trauma Systems Foundation (PTSF) for the purposes of evaluating the trauma system and identifying opportunities to improve performance of trauma centers and the trauma system overall.

The project was accomplished by developing a risk-adjusted mortality model for trauma centers in PA, and then using the model to estimate the relative performance of each trauma center. Relative performance was defined using observed-to-expected mortality ratios (O/E ratio). The O/E ratio is calculated by dividing the rate of mortality by the rate of expected mortality at a given trauma center. The expected mortality at trauma centers is calculated after accounting for differences in patients treated at different trauma center (i.e., case-mix adjustment). To accomplish this, the O/E ratio is estimated by developing a statistical model that uses patient demographic and clinical characteristics to estimate the probability of death for each patient. The resulting probabilities of death are used to determine, for each trauma center, the number of patient deaths that were expected to have occurred given the patients' underlying medical conditions and the severity of the patients' injuries.

In 2006, the American College of Surgeons (ACS) Committee on Trauma launched the Trauma Quality Improvement Program (TQIP) to study the variability in outcomes between trauma centers in the United States and Canada. The primary goal of TQIP is to improve the quality of trauma care through outcomes-based, risk-adjusted benchmarking of trauma centers and feedback reports.[1,2] An article that was recently published by Newgard et al. describes the methodology, data processing, data quality, statistical analysis, and analytic rationale for the method that TQIP recommends for use in efforts to develop risk-adjusted models to enable comparisons of trauma center performance.[3] The reader is referred to that article for additional background and details of the TQIP-recommended method for developing a risk-adjustment morality model. Using data elements available in the PTOS database our group replicated the methods of by Newgard et al. as closely as possible to develop the analysis and results reported here. Both the fundamental methods we used as well as how they differed from the model of Newgard et al. are described.

This document reports the results of the analysis that developed a risk-adjusted mortality model for the Pennsylvania Trauma Systems Foundation (PTOS-RAM model). First, we present the overall model that was applied to the entire patient sample. Subsequent sections present the variants of the model that separately investigated mortality for patients in subcategories of interest: blunt, penetrating, multisystem, isolated head injury, and geriatric.

SETTING AND DATA

The PTSF provided data from the Pennsylvania Trauma Outcomes Study (PTOS). The PTOS data used in the project included all records for patients age 16 years and older with trauma or

penetrating trauma who were treated at trauma centers in PA during 2011-2013. Pediatric trauma centers and trauma centers not accredited for the full study period were excluded. Trauma centers were indicated using an alias identification number so that the identity of a given trauma center was not revealed.

PATIENT AND TRAUMA CENTER CHARACTERISTICS

The study data included a total of 100,278 patients treated at 27 trauma centers. Characteristics of the patients are reported in Table 1. The amount of missing data was minimal and ranged from 0% missing on age, sex, discharge status (died, survived), and transfer status, to 5.8% missing on injury mechanism.

Table 1. Characteristics of 100,278 patients treated at 27 trauma centers in PA, 2011-2013.

Characteristic	Mean (SD) or %	% missing
Age, mean (SD)	55.5 (23.2)	0
Female, %	40.6	0
Died, %	4.9	0
Mechanism		7.8
Pedestrian/pedal, %	4.6	.
Motor vehicle occupant, %	23.5	.
Motorcyclist, %	5.2	.
Fall, %	54.0	.
Struck by / against, %	2.8	.
Firearm, %	4.8	.
Cut / pierce, %	2.6	.
Other, %	2.4	.
Transfer patient, %	30.6	0
Systolic BP, admission, mean (SD)	139.7 (31.1)	0.49
GCS, mean (SD)	5.62 (1.24)	5.75
Pulse, mean (SD)	86.8 (21.0)	0.32
Single worst injury (SWI), mean (SD)	0.06 (0.16)	
ISS, mean (SD)	10.3 (8.6)	0.95
AIS, head, mean (SD)*	2.5 (1.3)	0
Lowest AIS, mean (SD)	0 (0.0)	0
Heart disease, %	18.5	2.33
Cancer, %	1.6	2.33
Liver disease, %	1.1	2.33
Hypertension, %	42.9	2.33

Diabetes, %	16.1	2.33
Bleeding disorder, %	14.5	2.33
Arrest SBP, %	1.2	0.49
Shock, %	3.3	0.49
Multisystem, %	7.2	0
Isolated head, %	17.4	0
Admitted to ICU, %	38.0	0

Arrest SBP: systolic blood pressure in ED <= 40 mmHg.
Shock SBP: systolic blood pressure in ED < 90 mmHg.
Multisystem: AIS >= 3 on 2 or more AIS body regions.
Isolated head injury: AIS head >= 3 and AIS <3 on other AIS body regions.
* Calculated among patients with head AIS>=1.

Characteristics of patients treated at the 27 trauma centers are reported in Table 2. The second column shows that the number of patients treated ranged from a low of 1,756 to a high of 12,825 patients. The number of deaths ranged from 42 to 527 deaths and the percent mortality ranged from 2.2% to 9.5%. The percent of patients treated for blunt (as opposed to penetrating) trauma ranged from 78.1% to 97.6%.

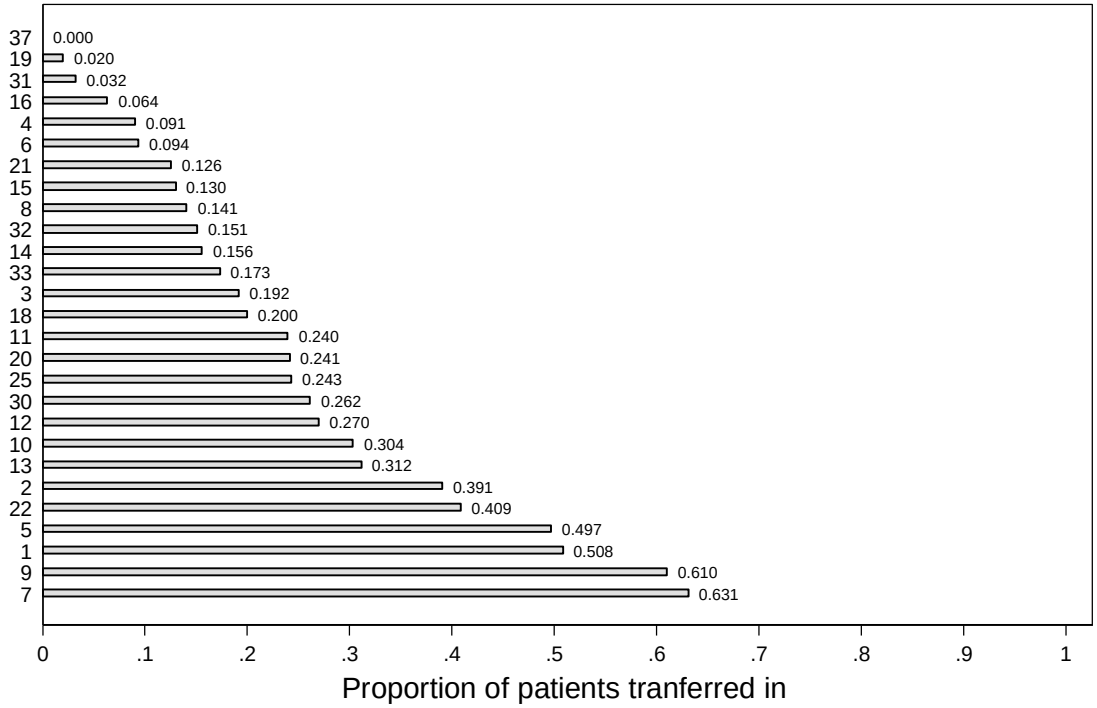
Table 2. Characteristics of patients (n=100,278) by trauma center (n=27).

ID	Patients	Deaths	Mortality, %	Age, mean (SD)	Female, %	Blunt, %	ISS, mean (SD)
1	4317	204	4.7%	55.0 (23.03)	40.7%	96.0%	12.4 (8.86)
2	7331	298	4.1%	59.9 (23.74)	46.4%	95.7%	10.5 (8.30)
3	3856	191	5.0%	57.7 (23.32)	44.4%	93.3%	9.9 (8.30)
4	2008	143	7.1%	47.5 (21.37)	31.5%	88.2%	11.0 (9.96)
5	7087	351	5.0%	56.0 (22.51)	41.1%	94.0%	10.2 (7.79)
6	2318	221	9.5%	49.8 (22.46)	33.6%	78.9%	11.2 (10.43)
7	3449	155	4.5%	55.6 (22.88)	37.0%	88.8%	10.3 (9.00)
8	3718	187	5.0%	55.3 (22.77)	42.3%	91.6%	9.3 (7.78)
9	12825	527	4.1%	53.7 (22.46)	38.5%	92.3%	8.5 (7.05)
10	3810	345	9.1%	44.5 (20.92)	24.5%	72.9%	11.5 (11.10)
11	4133	314	7.6%	45.6 (21.17)	28.5%	78.1%	12.4 (11.53)
12	3611	184	5.1%	51.8 (23.06)	37.1%	92.2%	14.7 (11.08)
13	1756	65	3.7%	53.5 (22.32)	39.0%	96.5%	10.1 (7.89)
14	3293	144	4.4%	55.5 (22.90)	42.3%	93.2%	9.0 (7.28)
15	2629	117	4.5%	66.9 (22.05)	54.3%	97.6%	9.3 (7.65)
16	3362	156	4.6%	60.0 (23.43)	46.0%	95.9%	11.6 (9.19)
18	3185	120	3.8%	54.6 (23.20)	39.2%	89.4%	8.9 (7.99)
19	1916	56	2.9%	64.7 (22.39)	52.8%	97.2%	10.2 (8.30)
20	2966	139	4.7%	62.1 (22.91)	47.3%	97.5%	10.8 (8.67)
21	1877	100	5.3%	55.0 (22.51)	42.5%	95.6%	10.2 (7.73)
22	3605	174	4.8%	51.7 (22.21)	36.9%	92.7%	10.7 (8.07)
25	3995	159	4.0%	59.7 (23.50)	45.5%	95.6%	9.9 (8.01)
30	2715	142	5.2%	51.5 (22.61)	36.9%	96.1%	11.0 (8.16)
31	3822	168	4.4%	62.0 (23.37)	44.8%	94.5%	10.8 (8.28)
32	2086	117	5.6%	55.7 (22.83)	41.9%	93.1%	10.3 (7.32)
33	2770	76	2.7%	58.8 (23.38)	44.4%	96.2%	10.4 (8.17)
37	1838	42	2.3%	59.4 (23.04)	47.3%	97.2%	8.2 (6.19)

ID is an alias number.

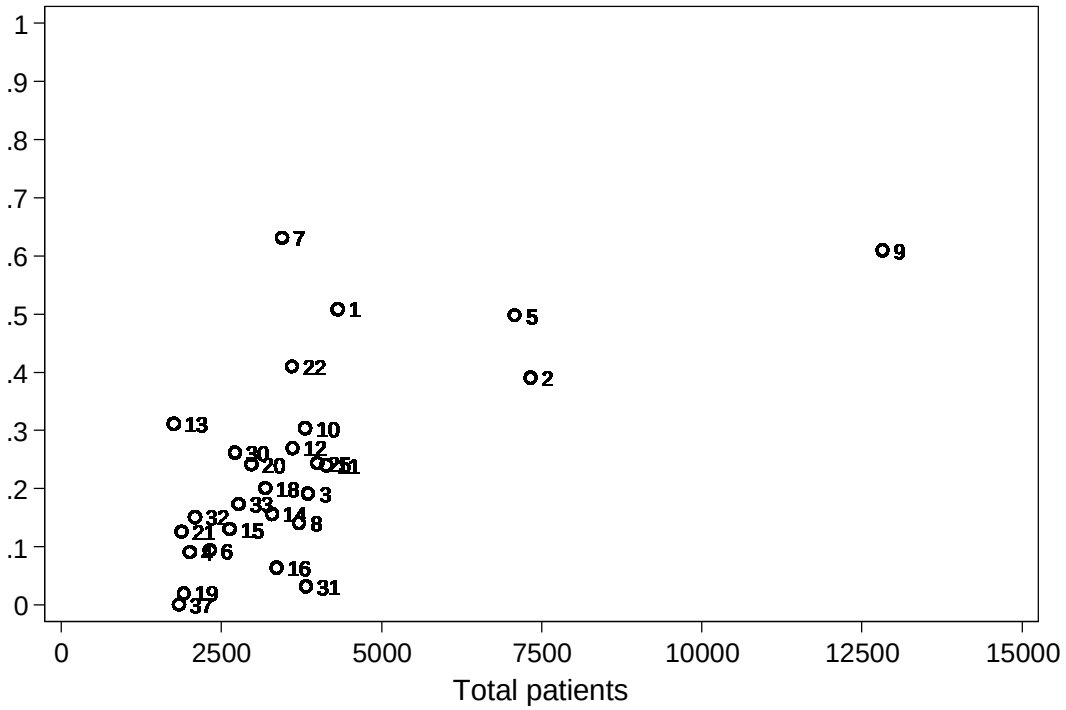
The proportion of transfer patients varied considerably across the 27 trauma centers (Figure 1). The proportion of transfer patients at centers ranged from 0% to 63%.

Figure 1. Proportion of patients at each trauma center who were transfer patients.



There was a systematic relationship between trauma center volume and proportion of transfer patients (Figure 2). In general, the proportion of transfer patients was higher in higher volume trauma centers (correlation coefficient = 0.74). The three trauma centers with > 5000 patients over the study period had between 40% and 60% transfer patients. Proportions of transfer patients were highly variable at centers with < 5000 patients ranging from 0% to 63% (shown at the left side of Figure 2).

Figure 2. Scatterplot comparing the proportion of transfer patients to the number of patients treated at the 27 trauma centers.



LOGISTIC REGRESSION STATISTICAL MODEL

All analyses were conducted using Stata 13.1 (College Station, TX). The statistical model was developed using logistic regression. The outcome of the statistical model was discharge status (died versus lived). A total of 16 variables were modeled as the independent, adjustor variables. One variable (lowest AIS) did not contribute statistically to prediction and was excluded. The variables (reported below) were selected to match, to the extent possible, the variables used by Newgard et al. in the final TQIP statistically model.

Given the presence of missing data, multiple imputation was conducted to impute values for values that were missing. The rationale for doing so was to avoid bias and imprecision that could result from using another method (e.g., complete case analysis, listwise deletion).[4] We used the multivariable normal regression method, which has been shown to be flexible and appropriate for use in continuous, dichotomous, and categorical data. Prior to imputation, continuous variables were examined for normality. Variables that were not normally distributed were log-transformed to approximate a normal distribution, then imputed, and afterward exponentiated to retain their original distribution.

A total of 10 new datasets in which missing values were replaced with imputed values were generated. Those 10 datasets were modelled simultaneously using logistic regression that pooled the effect estimates and adjusted the standard errors accordingly. First, a logistic regression model (Model 1) was developed that treated continuous variables in their original continuous form. Second, a logistic regression model (Model 2) was developed that treated the continuous variable as categorical (divided into seven groups of equal size). This was conducted to allow for a relationship between a predictor variable and the outcome that operated in a nonlinear fashion. Conventional diagnostics were conducted, including bivariate correlations and variance inflation factors to identify the presence of multicollinearily and plots of deviance versus leverage values to identify outlier observations.[5] Goodness of fit statistics, described below, identified that Model 2 was superior.

The final logistic regression model was used to generate the risk-adjusted predicted probability of death for each patient in the original sample of 100,278 patients. Those values ranged from 0 to 1, with higher values indicating a higher predicted probability that a given patient had died.

LOGISTIC REGRESSION RESULTS

Results of the final logistic regression model (Model 2) are reported in Table 3. As shown, the model included a total of 16 predictor variables. The effect estimates, which are odds ratios, can be interpreted as relative risks of death and can range from 0 to infinity. An odds ratio of 1 is the null value; values less than one indicate a protective effect and values greater than one indicate an increased risk of death.

Compared to the youngest patients, older patients were generally more likely to die, with patients in the oldest group being 10.98 times more likely than the youngest patients to die. Compared to patients treated for a pedestrian/pedal injury, motor vehicle occupant patients, fall patients, and struck by/against patients were less likely to die. Compared to pedestrian/pedal injury, patients injured with firearms (gunshot trauma) were more likely to die. Transfer patients were less likely to die than non-transfer patients.

Table 3. Results of the PTOS-RAM to estimate the risk of death based on patient demographic and clinical characteristics (n = 100,278).

		Odds ratio	Std. Err.	t	P-value	95% Conf. Interval	
Age group (years, mean)							
(20.2)	1 (youngest)	*-ref-*					
(30.5)	2	1.04	0.10	0.77	0.629	0.898	1.280
(44.5)	3	1.34	0.12	2.95	0.003	1.098	1.590
(55.4)	4	2.16	0.21	8.11	<0.001	1.805	2.631
(66.3)	5	4.35	0.41	15.01	<0.001	3.521	5.140
(78.5)	6	7.75	0.76	20.98	<0.001	6.420	9.419
(88.8)	7 (oldest)	10.98	1.13	24.17	<0.001	9.237	13.676

Mechanism							
	Pedestrian/pedal	*-ref-*					
	MV occupant	0.79	0.05	-3.68	<0.001	0.696	0.895
	Motorcyclist	0.91	0.09	-1.05	0.321	0.745	1.093
	Fall	0.69	0.04	-5.76	<0.001	0.618	0.789
	Struck by / against	0.39	0.07	-5.31	<0.001	0.267	0.544
	Firearm	4.26	0.34	17.96	<0.001	3.598	4.923
	Cut / pierce	0.81	0.12	-1.32	0.186	0.606	1.102
	Other	0.69	0.09	-2.81	0.005	0.560	0.902
Transfer patient		0.83	0.04	-4.30	<0.001	0.749	0.898
Systolic BP (admission, mean)							
(91.3)	1 (lowest)	*-ref-*					
(119.7)	2	0.39	0.03	-12.55	<0.001	0.332	0.447
(130.1)	3	0.40	0.03	-12.17	<0.001	0.341	0.460
(138.8)	4	0.34	0.03	-14.03	<0.001	0.294	0.396
(148.8)	5	0.35	0.03	-14.41	<0.001	0.303	0.403
(160.2)	6	0.32	0.02	-15.84	<0.001	0.273	0.364
(186.7)	7 (highest)	0.38	0.02	-15.11	<0.001	0.336	0.432
GCS, motor score							
	1 (lowest)	*-ref-*					
	2	3.36	0.48	8.49	<0.001	2.539	4.442
	3	1.48	0.20	2.94	0.003	1.139	1.915
	4	0.67	0.06	-4.54	<0.001	0.562	0.795
	5	0.31	0.02	-15.90	<0.001	0.272	0.362
	6 (highest)	0.07	0.00	-51.98	<0.001	0.063	0.077
Pulse rate (mean)							
(56.0)	1 (lowest)	*-ref-*					
(71.7)	2	0.76	0.06	-3.64	<0.001	0.656	0.881
(78.5)	3	0.67	0.05	-4.95	<0.001	0.572	0.786
(85.1)	4	0.82	0.06	-2.62	0.008	0.711	0.952
(91.9)	5	0.77	0.06	-3.29	<0.001	0.652	0.897
(100.4)	6	0.87	0.07	-1.83	0.040	0.753	1.009
(119.9)	7 (highest)	1.33	0.09	4.32	<0.001	1.168	1.510
Single worst injury							
	1 (lowest)	*-ref-*					
	2	1.65	0.27	3.11	0.001	1.204	2.306
	3	2.06	0.31	4.39	<0.001	1.454	2.678
	4	2.37	0.32	5.92	<0.001	1.736	2.996
	5	2.90	0.43	7.54	<0.001	2.235	3.953
	6	3.69	0.52	9.26	<0.001	2.799	4.886

7 (highest)	7.13	0.99	14.15	<0.001	5.421	9.379
Head AIS						
No head injury	*-ref-*					
1 (lowest)	1.04	0.08	0.57	0.529	0.899	1.213
2	0.81	0.08	-2.27	0.022	0.674	0.971
3	1.42	0.08	5.86	<0.001	1.261	1.592
4	1.34	0.11	3.69	<0.001	1.147	1.565
5	5.99	0.37	29.08	<0.001	5.306	6.754
6 (highest)	36.67	21.965	6.110	<0.001	11.640	118.382
Comorbidities						
Heart disease	1.46	0.08	7.14	<0.001	1.317	1.623
Cancer	1.74	0.19	4.91	<0.001	1.383	2.129
Liver disease	4.00	0.51	10.71	<0.001	3.094	5.132
Hypertension	0.93	0.05	-1.61	0.107	0.833	1.018
Diabetes	1.08	0.06	1.96	0.051	1.000	1.240
Bleeding disorder	1.51	0.08	7.52	<0.001	1.346	1.660
Arrest SBP	46.13	6.67	26.21	<0.001	34.376	60.899
Constant	0.05	0.01	-18.46	<0.001	0.034	0.065

CI: confidence interval.

MODEL FIT AND CALIBRATION

It is critical that the estimated logistic regression model effectively represents the data and serves as a valid tool for risk adjustment. This section evaluates the discrimination and calibration of the risk-adjusted model.

Discrimination

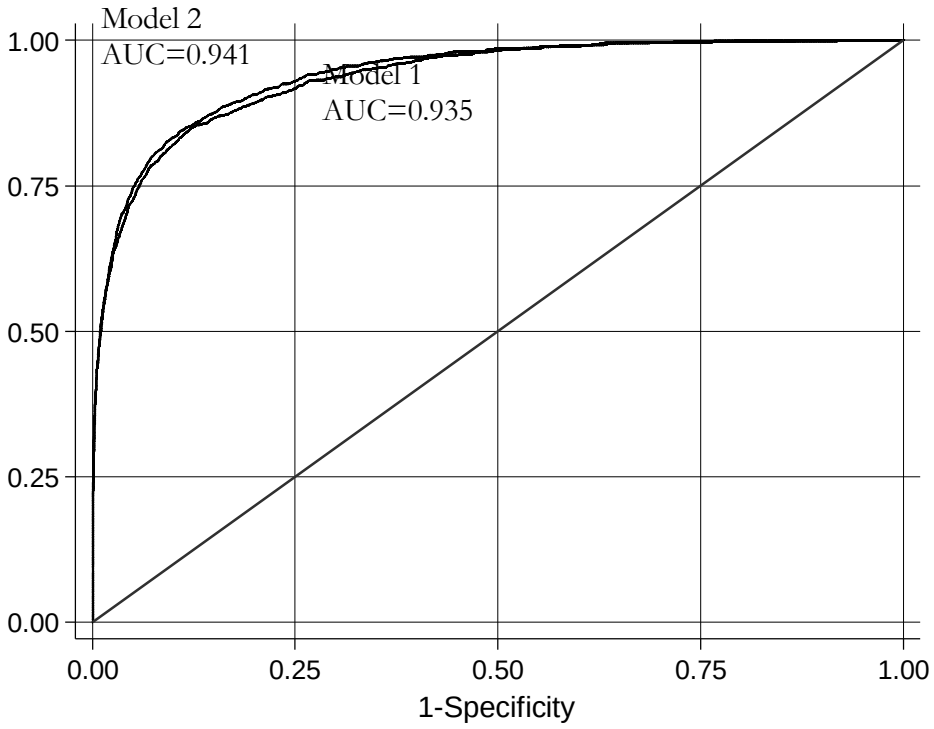
A statistical model that distinguishes well between patients who die and patients who survive is said to have good discrimination. A commonly used measure of discrimination is the c-statistic (or c-index), also referred to as the area under (AUC) the receiver operating characteristic (ROC) curve. The c-statistic ranges from 0 to 1, with better discrimination indicated by higher values. A c-statistic of 0.5 indicates that the ability of the model to discriminate between patients who die and patients who survive is no better than chance. Values that deviate away from 0.5 indicate that the model performs better than chance. The specific value of the c-statistic has a meaningful interpretation. If the c-statistic is 0.7, this suggests that if one patient who died and one patient who survived are drawn at random from the data, there is a 70% chance the patient died would have been assigned a higher predicted probability of death.

A graph of the area under the ROC curve also provides useful information. A model that discriminates well will produce an area under the ROC curve that has a high arch and approaches the upper left corner of the graph.

Figure 3 displays the graph of the area under the ROC curve for the final statistical model (Model 2), and indicates that the c-statistic for the final model was 0.941 (95% CI = 0.935 to 0.946). As noted above, in this final model continuous variables were treated as categorical to allow for nonlinear relationships. Figure 3 also reports that the model that included continuous variables in their original continuous form (Model 1) had an area under the curve of 0.935 (95% CI = 0.929 to 0.941). Despite being nearly identical, the AUC for Model 2 was larger than the AUC for Model 1 when evaluated with a statistical test (chi-square (1)=12.56, p=0.0004) and thus is the preferred model based on this criterion.

In comparison, the c-statistic reported recently in by Newgard et al. in their model to demonstrate the TQIP method of risk adjustment was 0.90. Thus the PTOS-RAM model appears to discriminate as well, or better, than the model used by the TQIP program that computed risk-adjusted outcomes for demonstration purposes.

Figure 3. Graph of area under the receiver operating characteristic curve to evaluate discrimination of the PTOS-RAM model. Model 1: continuous variables coded as continuous. Model 2 (final model): continuous variables coded as categorical.



Calibration

Calibration refers to the ability of a model to match predicted and observed death rates across the span of injury severity that was observed in the data. The span of injury severity can be expressed by arranging the patients from those with the least to those with the most severe injuries, and dividing the patients into ten groups of equal size. A model demonstrates good calibration if the number of deaths predicted by the model in each decile closely matches the number of actual deaths that were observed in each decile.

The Hosmer-Lemeshow chi-square statistic is a common measure of calibration that compares observed and predicted outcomes over deciles of risk. The Hosmer-Lemeshow statistic for Model 1 model was 70.08 (df=10, p-value<0.0001) and the Hosmer-Lemeshow statistics for Model 2 was 55.43 (df=10, p-value<0.0001) (Table 4). A smaller test statistic, and larger p-value, indicates better calibration, and a p-value > 0.05 lets the analyst reject the null hypothesis of no difference between actual and predicted deaths. Neither Model 1 nor Model 2 met this

criterion. This is not a concern, however, given that it is common for the Hosmer-Lemeshow test to fail when used in a large sample.

An additional measure of goodness-of-fit is the Akaike information criterion (AIC), which enables comparisons of different statistical models and indicated and where smaller values indicate a better fit. As reported in Table 4, the AIC value was 20695.36 for Model 1 and 20084.36 for Model 2.

Based on these three criteria – the Hosmer-Lemeshow statistic, the c-statistic, and the AIC – Model 2 fit the data better than did Model 1. Thus Model 2 was selected as our final model. In the reminder of the report Model 2 is referred to as the PTOS risk-adjusted mortality (PTOS-RAM) model.

Table 4. Comparison of multivariable logistic regression models for risk-adjustment of PTSF patients (n=100,278).

	Model 1	Model 2
No. of variables	16	16
Coding of continuous variables	Continuous	Categorical
Model performance		
Hosmer-Lemeshow goodness of fit	70.08 (10), p<0.0001	55.43 (10), p<0.0001
C-statistic	0.935	0.941
AIC	20695.36	20084.36

Model 1 and Model 2 included age*, injury mechanism, transfer status, SBP*, GCS motor score, pulse*, SWI (single worst injury based on ICD-9 injury codes)*, head AIS, lowest AIS, heart disease, cancer, liver disease, hypertension, bleeding disorder, diabetes, and arrest SBP (emergency department SBP<=40 mmHg). Lowest AIS was not statistically significant and was dropped from the model.
Asterisk (*) indicates variables that were used in their original continuous form in Model 1 and that were recoded into categorical variables (each divided into seven equal-sized groups) and used as categorical in Model 2.
Variables used in TQIP final model that were not included in the PTSF model: impaired sensorium, functional dependence, dialysis, and peripheral vascular disease.
Variables not used in TQIP final model that were included in the PTSF model: diabetes.
For the Hosmer-Lemeshow goodness of fit statistic, a p-value > 0.05 indicates a better fit.
The c-statistic ranges from 0 to 1, with higher values indicating better model discrimination.
When comparing the fit of multiple models, a lower Akaike information criterion (AIC) value indicates better model fit.

We conducted additional diagnostics on the results of the final PTOS-RAM model to further evaluate how well it was calibrated to predict mortality among the 100,278 patients overall. There are multiple common approaches to assess calibration that involve inspecting the difference between the actual number of deaths that occurred in groups of patients divided according to decile or risk of death. First we report on results using true deciles, and second we present results using 10 groups of equal size. The implication for multiple approaches will emerge in the descriptions of the methods.

Inspecting the calibration of the model based on the difference between the actual number of deaths that occurred in each (true) decile of risk and the number of deaths that were predicted in each (true) decile of risk can be seen in Table 5. Of the 100,278 patients in the dataset, a total of 4,928 patients died and as expected the number of patients that died was higher in higher deciles of risk. A total of 4,928 deaths were also predicted, and the number of deaths that were predicted was higher in higher deciles of risk. The PTOS-RAM model consistently under-predicted the number of deaths that occurred in the lowest seven deciles of risk, however, and over-predicted the number of deaths that occurred in the higher deciles of risk. This reveals that the model was calibrated well overall but that the predicted probability of death was miss-specified to a degree as a function of injury severity.

Table 5. Calibration of PTOS-RAM within patient subgroups of interest.

Subgroup	C-statistic	95% CI
Blunt	.932	.925, .938
Penetrating	.980	.974, .986
Multisystem	.879	.862, .896
Isolated head	.918	.907, .927
Geriatric	.883	.869, .896

We also evaluated how well the PTOS-RAM was calibrated within patient subgroups of interest. Table 5 shows the c-statistic values that range from 0.879 among multisystem patients to 0.980 among penetrating trauma patients. Thus, variability was observed, but calibration was good among each of these subgroups.

Table 6. Actual and predicted deaths by decile of risk as a measure of calibration of the PTOS-RAM model.

Decile	Patients	Minimum predicted risk	Maximum predicted risk	Predicted deaths	Actual deaths	Difference
1	10,026	0.000	0.001	9	1	8
2	10,029	0.001	0.002	17	4	13
3	10,025	0.002	0.003	27	10	17
4	10,030	0.003	0.005	42	26	16
5	10,025	0.005	0.008	68	43	25
6	10,031	0.008	0.013	107	94	13
7	10,028	0.013	0.020	162	131	31
8	10,028	0.020	0.032	254	271	-17
9	10,028	0.032	0.080	490	541	-51
10	10,028	0.080	1.000	3,719	3,774	-55
Total	100,278			4,895	4,895	

Figure 4 was generated using the information in Table 6 and reports additional information that is used to evaluate the calibration of the PTOS-RAM model. The graph plots the observed versus the predicted morality (percent of patients who died) in each of the 10 deciles of risk, and a 45-degree line that represents perfect calibration. The closer the predicted values are to the 45-degree line, the better the calibration. It is visually evident that, overall, the predictions track observed mortality well.

Figure 4. Graph observed versus predicted mortality plotted in deciles of predicted risk as a measure of calibration of the PTOS-RAM model in 100,278 patients.

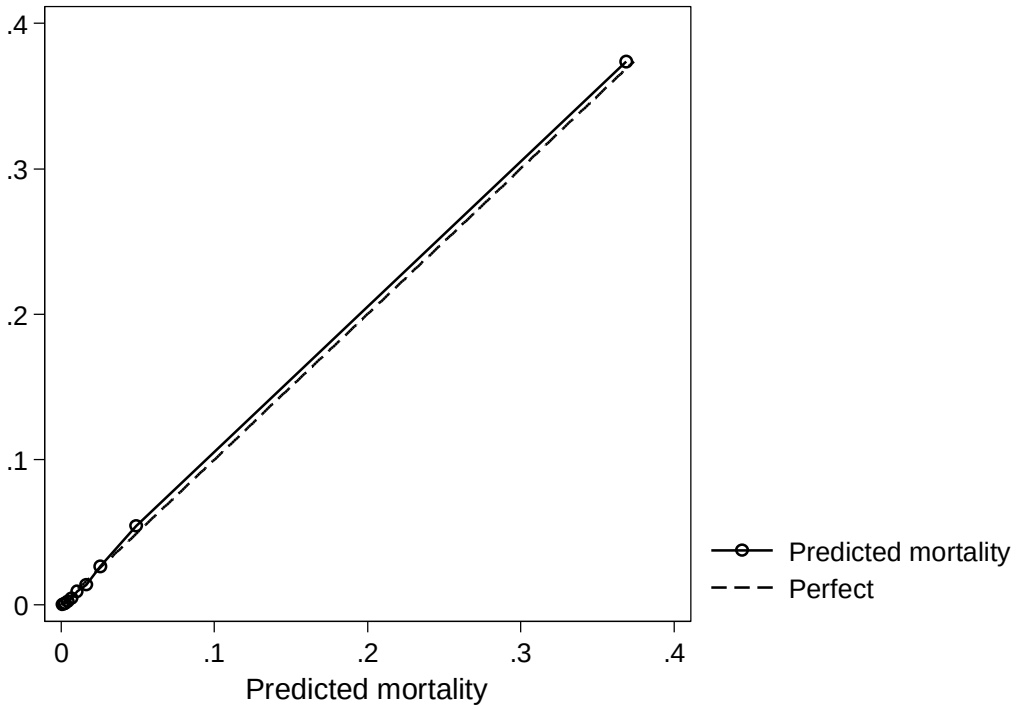


Table 6 also reports the minimum and maximum predicted risk of death for patients within each decile. Both minimum values and maximum values increased consistently when considering the deciles from lowest to highest levels of risk, which is expected given that the deciles were constructed in order of increasing predicted probability of death. Note that these columns provide useful information, by revealing that the predicted probability of death is a maximum of 0.0810 (i.e., 8.1%) in the ninth decile. This is an indication of the fact that death was relatively rare in the total sample: 4,895 of 100,278 = 4.9% of patients died. As a result, splitting the sample into deciles of risk creates nine deciles where the probability of death was very low and a tenth decile where the probability was considerably higher and spanned a very wide range of predicted probabilities.

A practical implication is that a better understanding of model fit would emerge by dividing the sample not into deciles of predicted probabilities of death but instead into 10 groups that represent patients with a probability of death <=10%, 11-20%, 21-30%, …, 91-100%. Results based on this approach are reported in Table 7. Arranged in this way, the number of patients in each group varies widely from a minimum 397 patients to a maximum of 91,558 patients. The final column of the table indicates that the PTOS-RAM model over-predicted the number of

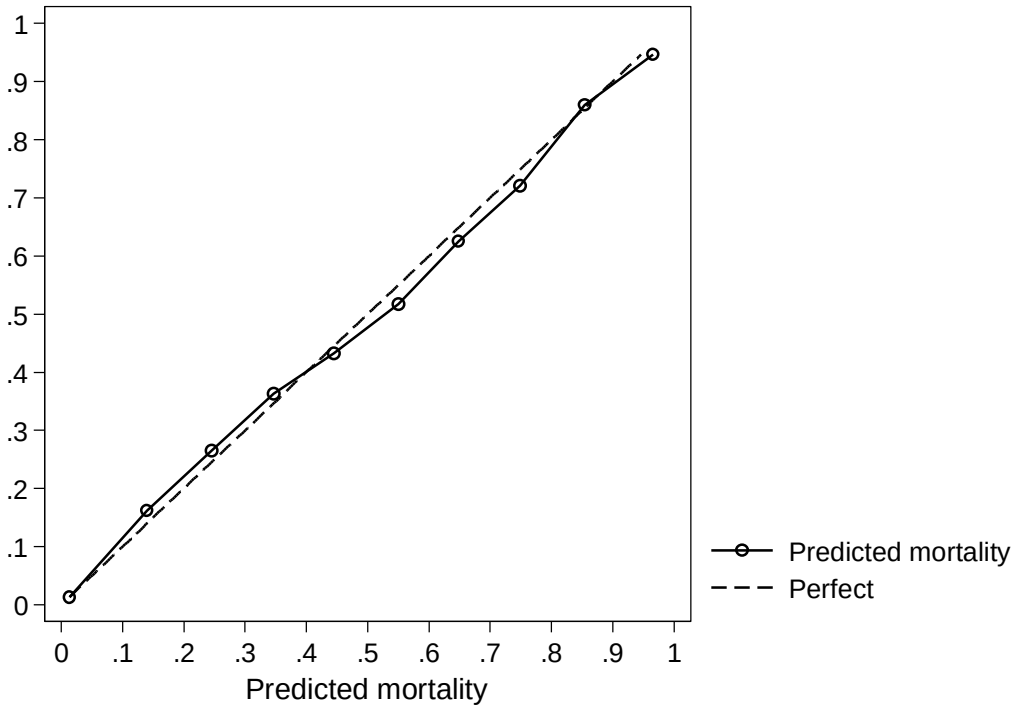
deaths in the lowest risk group by 45 patients, under-predicted the number of deaths in the next three groups, and over-predicted the number of deaths in the remaining groups except for the second highest group.

Table 7. Actual and predicted deaths within 10 groups representing the range of predicted probabilities of death as a measure of calibration of the PTOS-RAM model.

Group	Patients	Minimum predicted risk	Maximum predicted risk	Predicted deaths	Actual deaths	Difference
1	91,558	0.000	0.100	1,294	1,251	43
2	3,169	0.100	0.200	443	509	-66
3	1,279	0.200	0.300	315	344	-29
4	820	0.300	0.400	285	299	-14
5	599	0.400	0.500	266	258	8
6	469	0.500	0.600	258	241	17
7	418	0.600	0.700	272	261	11
8	397	0.700	0.800	297	288	9
9	454	0.800	0.900	388	386	2
10	1,115	0.900	1.000	1,078	1,058	20
Total	100,278			4,895	4,895	

A plot of observed versus expected probability of mortality that is based on the information in Table 7 is shown in Figure 5. It becomes apparent that the advantage of this approach to creating patient groups is that, when graphed, the results reveal with finer granularity the extent to which the risk-adjusted mortality model was calibrated across the span of probabilities of death. Figure 5 reveals that the model predicts risk quite well across the span of probabilities of death.

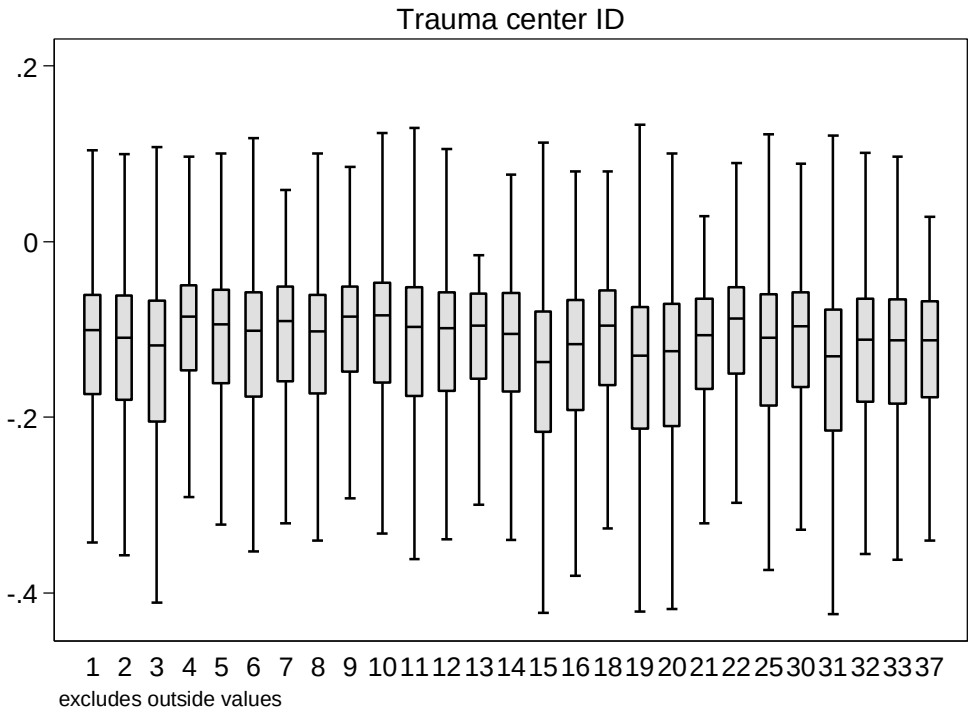
Figure 5. Graph observed versus predicted mortality plotted in ten groups of risk equally spaced from 0% to 100% predicted risk as a measure of calibration of the PTOS-RAM model in 100,278 patients.



Residuals

A residual for each patient can be generated from the logistic regression model. Patients who lived are coded '0' and patients who died are coded '1', and the predicted likelihood that a given patient died is represented as a decimal value ranging from 0 to 1, where higher values indicate a higher predicted likelihood that the patient died given their clinical characteristics and injury severity. The residual is the difference between the predicted likelihood of death value and the patient's true value (i.e., 0 or 1). By graphing boxplots of the residuals of the PTOS-RAM model for each trauma center in Figure 6, it can be seen that the median residual was consistent across the 27 trauma centers. (In a boxplot, the horizontal line in each box indicates the median and the lower and upper ends of the box indicated the 25% and 75% percentiles, respectively). This is an indication that the statistical model generally over-predicted expected mortality for patients. This is not evidence that the model fit poorly, however; it is not surprising given that death in the total sample was a rare event. Ultimately the model diagnostics that were conducted did not find systematic errors in specification as a function of trauma center characteristics.

Figure 6. Boxplot of results from the PTOS-RAM as an indication of model fit for patients at each of the 27 trauma centers.



TRAUMA CENTER OBSERVED-TO-EXPECTED MORTALITY

Having developed a statistical model that was well calibrated and fit the data well, the final step was to determine how the observed mortality at each trauma center compared to the mortality that was expected given the characteristics and injury severity of its patients. Dividing the observed death rate at a hospital by the expected death rate at the hospital produces the O/E ratio. If a hospital had more deaths than expected, the O/E ratio will be greater than 1. If the hospital had fewer deaths than expected, the O/E ratio will be less than 1. O/E ratios methodology provides an easy way for facilities to be benchmarked against each other and has been widely used by the trauma community for this purpose.

The end goal of generating risk-adjusted O/E ratios using the PTOS data is to provide readily interpretable and informative comparisons of the performance of trauma centers in Pennsylvania. A rank plot (also called caterpillar plots) is a familiar format used traditionally to enable comparative "ranking" of hospitals." As Newgard et al. point out, these plots have important limitations. In particular, a rank ordering is not necessarily meaningful given that two trauma centers may be ranked far apart but yet may not be statistically different in how they performed. Also, caterpillar plots do not indicate trauma center volume and thus cannot be used to compare outcomes among centers with similar volumes. Caterpillar plots also make it difficult to identify trauma centers that qualify as statistical outliers. For these and other reasons, TQIP recommends using funnel plots to report on trauma center performance. Funnel plots allow trauma center volume to be directly assessed, and enable improved visual assessment of outlier hospitals (high and low), elimination of non-meaningful hospital rankings, and easier identification of hospitals close to outlier status (e.g., early recognition of quality issues that can prompt behavior change, even if not yet statistically significant).[3,6,7]

In order to allow for trauma center volume to be directly assessed using funnel plots, each hospital's performance is plotted against its effective sample size (i.e., for a Poisson distribution, this is the number of events). When graphing observed-to-expected mortality it is conventional to plot the O/E ratio for each hospital against expected mortality.[8] Unlike a caterpillar plot, where confidence intervals indicate uncertainty for each observed hospital O/E ratio, in a funnel plot a single envelope of uncertainty is drawn from the expected line. This line can be thought of as a threshold for "alert" and hospitals that lie outside the expected line (e.g., 95% confidence interval for the entire sample of hospitals) can be considered further for characteristics that may be associated with their disproportionately above- or below-expected mortality.[9]

Figure 7 presents a caterpillar plot and a funnel plot to report the observed-to-expected mortality ratios for the 27 trauma centers in PA based on the PTOS-RAM analysis of all 100,278 patients. The funnel plot identifies seven trauma centers that have outlier status, two of which have a statistically higher than expected mortality rate (center 6 and 9) and five of which have a lower than expected mortality rate (center 19, 25, 31, 33, and 37). Note, for example, that trauma center 9, which has a higher than expected mortality rate, had by a

considerable margin the highest number of expected deaths during the study period (approximately 460). Trauma center 6 had less than 200 expected deaths but also had a higher than expected mortality rate. Trauma centers 25 and 31 had approximately the same number of expected deaths (about 200) but had lower than expected mortality. Three other trauma centers were also were expected to have about 200 deaths but had observed-to-expected mortality rates within the bounds of what was expected. A large version of the funnel plot, which indicates each trauma center ID, is shown in the Appendix.

Figure 7. Visual depiction of risk-adjusted observed-to-expected mortality ratios for 27 trauma centers in Pennsylvania (based on analysis of 100,278 patients). Top. Rank (caterpillar) plot. Bottom. Funnel plot. Dashed lines denote 95% confidence limits.

Figure 7 Removed for Confidentiality

In addition to performing the analysis to examine overall risk-adjusted mortality at trauma centers in PA, analyses were conducted to examine risk-adjusted mortality at trauma centers within subgroups for blunt trauma, penetrating trauma, multisystem trauma, isolated head injury, and geriatric patients. Table 8 reports the incidence of blunt trauma, penetrating trauma, multisystem trauma, isolated head injury, geriatric patients, and also shock at each of the 27 trauma centers to report how common each of the injury types and conditions was at each trauma center.

Table 8. Incidence of blunt trauma, penetrating trauma, multisystem trauma, isolated head injury, geriatric patients, and shock at trauma centers in PA, 2011-2013.

ID	Patients	Blunt	Penetrating	Multisystem	Isolated head	Geriatric	Shock	ICU
1	4317	96.0%	4.0%	11.3%	18.9%	39.1%	2.6%	21.4%
2	7331	95.7%	4.3%	6.4%	18.9%	48.4%	1.9%	42.4%
3	3856	93.3%	6.7%	5.9%	19.3%	43.4%	3.7%	36.0%
4	2008	88.2%	11.8%	8.5%	16.0%	22.5%	5.1%	50.7%
5	7087	94.0%	6.0%	7.2%	20.3%	38.1%	3.2%	41.1%
6	2318	78.9%	21.1%	7.7%	17.7%	28.8%	6.8%	52.5%
7	3449	88.8%	11.2%	5.7%	22.2%	37.5%	2.1%	34.4%
8	3718	91.6%	8.4%	5.4%	15.2%	37.1%	3.1%	36.9%
9	12825	92.3%	7.7%	5.4%	16.9%	33.9%	2.3%	35.9%
10	3810	72.9%	27.1%	8.7%	18.9%	16.7%	9.0%	51.0%
11	4133	78.1%	21.9%	12.3%	14.6%	20.8%	4.9%	46.2%
12	3611	92.2%	7.8%	16.6%	15.6%	31.9%	3.8%	32.5%
13	1756	96.5%	3.5%	7.5%	10.9%	34.0%	2.9%	50.1%
14	3293	93.2%	6.8%	4.4%	17.2%	38.2%	3.2%	36.8%
15	2629	97.6%	2.4%	4.3%	17.8%	61.2%	3.0%	44.7%
16	3362	95.9%	4.1%	7.7%	20.2%	48.7%	2.7%	55.1%
18	3185	89.4%	10.6%	5.1%	15.9%	36.5%	2.6%	36.1%
19	1916	97.2%	2.8%	5.6%	17.4%	56.2%	2.5%	43.1%
20	2966	97.5%	2.5%	6.9%	17.7%	51.5%	2.7%	19.4%
21	1877	95.6%	4.4%	5.9%	15.7%	37.1%	3.7%	22.9%
22	3605	92.7%	7.3%	9.0%	17.2%	30.7%	3.4%	36.1%
25	3995	95.6%	4.4%	6.5%	18.9%	46.1%	1.9%	35.5%
30	2715	96.1%	3.9%	7.5%	19.9%	31.2%	3.9%	54.1%
31	3822	94.5%	5.5%	8.6%	15.8%	52.3%	4.7%	45.1%
32	2086	93.1%	6.9%	9.0%	13.0%	39.2%	4.3%	22.8%
33	2770	96.2%	3.8%	5.6%	19.4%	45.2%	2.6%	34.8%
37	1838	97.2%	2.8%	3.2%	10.4%	46.7%	3.1%	9.0%

Funnel plots reporting the results of the risk-adjusted mortality analyses with the subgroups of interest are presented below in Figure 8 (blunt trauma), Figure 9 (penetrating trauma), Figure 10 (multisystem trauma), Figure 11 (multisystem trauma), and Figure 12 (geriatric trauma). Large versions of each funnel plot are shown in the Appendix. In the funnel plots, O/E ratios are not plotted in a number of instances for specific trauma centers. This occurred when the O/E ratio could not be calculated due to a zero denominator.

Figure 8. Plots providing a visual depiction of risk-adjusted observed-to-expected mortality ratios for blunt trauma patients. A. Rank (caterpillar) plot. B. Funnel plot. 95% CIs that reach 4.0 are truncated.

Figure 8 Removed for Confidentiality

Figure 9. Plots providing a visual depiction of risk-adjusted observed-to-expected mortality ratios for penetrating trauma patients. A. Rank (caterpillar) plot. B. Funnel plot. 95% CIs that reach 4.0 are truncated.

Figure 9 Removed for Confidentiality

Figure 10. Plots providing a visual depiction of risk-adjusted observed-to-expected mortality ratios for multisystem trauma patients. A. Rank (caterpillar) plot. B. Funnel plot. 95% CIs that reach 4.0 are truncated.

Figure 10 Removed for Confidentiality

Figure 11. Plots providing a visual depiction of risk-adjusted observed-to-expected mortality ratios for isolated head trauma patients. A. Rank (caterpillar) plot. B. Funnel plot. 95% CIs that reach 4.0 are truncated.

Figure 11 Removed for Confidentiality

Figure 12. Plots providing a visual depiction of risk-adjusted observed-to-expected mortality ratios for geriatric trauma patients. A. Rank (caterpillar) plot. B. Funnel plot. 95% CIs that reach 4.0 are truncated.

Figure 12 Removed for Confidentiality

Table 9 summarizes the results reported in the funnel plots by listing the trauma centers that had an observed-to-expected mortality ratio that fell above or below the limit of expected threshold (i.e., 95% confidence interval) for all patients and for the subgroups of interest.

Table 9. Trauma centers with mortality observed-to-expected (O/E) ratio outside the 95% confidence interval in funnel plot analysis.

Table 9 Removed for Confidentiality

In the funnel plots shown above, it was not possible to list each trauma center by number given the large number of trauma centers and instances where points fell on top of one another. Thus we present Table 10 below, which lists the O/E ratio for each trauma center, presented in rank order from lowest to highest, for the analysis conducted on patients overall and for patients within subgroups. These are the same rankings that are used to order the trauma centers in each of the caterpillar plots. The table provides novel information in reporting the actual value of each O/E ratio. Also, each of the funnel plots shown above is shown in an

alternative version in the appendix, with a larger format so that each trauma center is shown by identification number.

Table 10. Mortality observed-to-expected (O/E) ratios for all 27 trauma centers presented in rank order, from lowest to highest, for all patients and for each subgroup.

Table 10 Removed for Confidentiality

SUMMARY

This effort produced a well-calibrated risk-adjusted mortality model for 2011-2013 PTOS data that generated estimates of the observed-to-expected mortality ratios (O/E ratio) for trauma centers in PA overall and for patient subgroups of interest. The overall model identified 2 trauma centers that had higher mortality than expected and 5 trauma centers that had lower mortality than expected. Variability was identified in the relative performance of trauma centers for the conditions of interested that were represented by the patient subgroups.

We were able to closely follow the guidelines for modeling risk-adjusted mortality recommended by the TQIP. Key differences in our method were as follows. Whereas the final TQIP model included 18 variables, our final model included 16 variables. Fifteen of those 16 variables were variables included in the TQIP final model and one variable, diabetes, was included in our model but not in the TQIP model. We included diabetes in an effort to proxy dialysis, which is a variable that TQIP included but that was not available in the PTOS data. The other variables included in TQIP but not included in our final model were impaired sensorium and functional dependence, but as our model performed very well in terms of calibration (c-statistic=0.941) and discrimination we do not believe that the absence of these variables negatively impacted our model.

Although the c-statistic (i.e., the area under the ROC curve) for the final model was very high, it is possible that a different regression modeling approach could have provided an even better fit to the data. For example, using a mixed-effects logistic regression model with random effects to represent patient subgroups would have allowed for instances where the relationships between fixed effects covariates and the outcome may have varied[3]. Despite this potential benefit, the basic approach to logistic regression modeling performs adequately and in addition is a well known technique that can be readily adopted. This facilitates a common modeling approach that allows for more direct comparisons between the results of different groups that develop risk-adjusted models.

While the TQIP guidance for risk-adjusted modeling lists steps for developing a model for a patient population overall, we took the additional step of developing O-E ratios for patient population subgroups. This step was not detailed in previous methodological descriptions, and it is quite possible that use of a mixed-effects model for these subgroups could improve the accuracy of subgroup-specific O-E ratios. We chose not to explore this option based on our

desire to be consistent with TQIP guidelines and on evidence that our final model was well-calibrated in the subgroups of interest. This is a topic that we would like to investigate in partnership with PTSF in future work to develop guidance for subgroup analysis and to evaluate the extent to which the results of a specific subgroup analysis match the subgroup results presented here.

We believe the results will be of use to the PTSF for evaluating model performance and performance of the trauma system overall. As one example, compare the results presented for the patients treated for blunt trauma (Figure8). The funnel plot (Figure 8) indicates considerable variability in mortality among patients treated for blunt trauma. Three of the 27 trauma centers had higher than expected mortality among blunt trauma patients. Also, five trauma centers had lower than expected mortality among blunt trauma patients. These findings may suggest that processes of care for blunt trauma management may account for the majority of variability in O/E mortality ratios.

Additional interesting findings emerged. For example, the subgroup analysis on penetrating trauma revealed that no trauma centers had higher than expected mortality and no trauma centers had lower than expected mortality in terms of the management of penetrating trauma. Alternatively, no trauma centers had lower than expected morality in the management of multisystem trauma but four trauma centers had higher than expected mortality in the management of multisystem trauma. One possible interpretation of these results is that training in and protocols for the management of penetrating trauma are uniformly robust, but best practices for the management of multisystem trauma may need further elucidation.

The results of the risk adjusted mortality analysis that are presented in this report should be used as a guide to help explore characteristics of trauma centers, patients, and other factors including geography that may influence trauma center performance.

REFERENCES

1. Hemmila MR, Nathens AB, Shafi S, et al. The Trauma Quality Improvement Program: Pilot Study and Initial Demonstration of Feasibility. *Journal of Trauma-Injury Infection and Critical Care.* Feb 2010;68(2):253-261.

2. Shafi S, Nathens AB, Cryer HG, et al. The Trauma Quality Improvement Program of the American College of Surgeons Committee on Trauma. *Journal of the American College of Surgeons.* Oct 2009;209(4):521-530.

3. Newgard CD, Fildes JJ, Wu L, et al. Methodology and Analytic Rationale for the American College of Surgeons Trauma Quality Improvement Program. *Journal of the American College of Surgeons.* 2013;216(1):147-157.

4. Allison PD. *Missing data.* Thousand Oaks, CA: Sage; 2002.

5. Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. *Am J Pub Health.* 1991;81(12):1630-1635.

6. Mullins RJ, VeumStone J, Hedges JR, et al. Influence of a statewide trauma system on location of hospitalization and outcome of injured patients. *Journal of Trauma-Injury Infection and Critical Care.* Apr 1996;40(4):536-546.
7. Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med.* Apr 30 2005;24(8):1185-1202.
8. Mohammed MA, Deeks JJ. In the Context of Performance Monitoring, the Caterpillar Plot Should Be Mothballed in Favor of the Funnel Plot. *The Annals of Thoracic Surgery.* 2008;86(1):348.
9. Spiegelhalter D. Funnel plots for institutional comparison. *Quality and Safety in Health Care.* December 1, 2002 2002;11(4):390-391.