IISE TRANSACTIONS ON HEALTHCARE SYSTEMS ENGINEERING
2017, VOL. 7, NO. 3, 181–191
https://doi.org/10.1080/24725579.2017.1325948

Taylor & Francis
Taylor & Francis Group

HEALTH INFORMATICS ON PROCESS IMPROVEMENT

Relative mortality analysis: A new tool to evaluate clinical performance in trauma centers

Nicholas J. Napoli[a], William Barnhardt[b], Madeline E. Kotoriy[c], Jeffrey S. Young[d], and Laura E. Barnes[a]

[a]Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA; [b]Emergency Services, University of Virginia Health System, Charlottesville, VA, USA; [c]Batten School of Leadership and Public Policy, University of Virginia, Charlottesville, VA, USA; [d]Department of Surgery, University of Virginia, Charlottesville, VA, USA

ABSTRACT

Improving trauma performance relies on outcome measures to target groups of patients with suboptimal outcomes. However, this is difficult when examining trauma data sets dominated by patients with high probability of survival (POS). The W-Score, a standard metric for evaluating trauma performance, disproportionately weights these populations and inaccurately represents effectiveness across the entire patient spectrum. We introduce the Relative Mortality Performance Trend (RMPT) and the Relative Mortality Metric (RMM), which provide valuable insight into trauma center performance at all levels of acuity and establish a more reliable metric for evaluating performance. We validate this method using data from a Level 1 trauma center over a 20-year period, where 89.39% of the patient population has a $POS > .90$. The RMPT groups patient populations by acuity levels, which allows us to identify changes in performance and isolate areas for improvement. We significantly outperformed the anticipated mortality with 95% confidence intervals across all POS ranges, except for the (0.799–0.901) and (0.967–0.970) ranges, which are targeted for improvement. The most significant improvements occurred for patients with $POS < 0.569$ between the 1994–1999 and 2003–2008 cohorts. Monte Carlo Simulations demonstrated that the RMM is consistently a more accurate metric than the W-Score when utilizing low sample sizes.

ARTICLE HISTORY
Received December 2016
Accepted April 2017

KEYWORDS
Trauma performance; relative mortality analysis; W-Score; Probability of Survival (POS); Trauma and Injury Severity Score (TRISS); mortality analysis; low sample size; quality; process improvement

1. Background

Since the publication of *Accidental Death and Disability* in 1966, care of traumatically injured patients in the United States has undergone a remarkable evolution (Biffl *et al.*, 2005). Care to the best of each physician's varying experience and ability has been replaced by standardized workflows, evidence-based medicine, and a streamlined continuum of care from the roadside to the operating room (Petrie *et al.*, 1996). All of these changes in workflows and operations procedures require evaluation to determine whether the change is advantageous, which in turn produces a metric.

When we measure performance using established metrics that redefine clinical operations and practices, we begin to start performing to suit the set metric; this is known as performativity. This construction of performativity is dangerous when the metric is not adequate or simply does not capture the true intention of the goal. Therefore, it is critical to ensure that these metrics are appropriate and unbiased.

Methods for evaluating performance, such as "morbidity and mortality conference," "preventable death studies," and "audit filters," are all subjective approaches in which the criteria is examined ad hoc without a systematic approach (Mock *et al.*, 2004). One of the first objective quantitative metrics, overall mortality count, is an insufficient measure of trauma care because of the lack of information that is provided regarding the patient's risk of death based on his or her physiological status and degree of injury.

As the progression of care has become more structured, quantitative measures such as the Trauma and Injury Severity Score (TRISS) (Boyd *et al.*, 1987) have been the prominent utility to evaluate trauma care (Osler et al., 2007). The TRISS method accounts for an extensive foundation of information regarding a patient's physiological and anatomical criteria (Hollis *et al.*, 1995; Schluter, 2011). Using a patient's injuries and initial vital signs, TRISS predicts the probability of survival (POS) of a patient based on and validated by national trauma registries (Schluter, 2011; Bouillon *et al.*, 1997). This adjusted risk of mortality can be thought as a patient's acuity level and therefore a measurement of the intensity of care required. The utilization of these TRISS derivative metrics goes beyond simply reporting performance for rankings, and is used for comparative analysis for validating process changes such as pre-hospital triage and other quality improvement changes (Mock *et al.*, 2004; Bouzat *et al.*, 2015). This article focuses on the current metrics and a new proposed metric based around TRISS for evaluating a trauma center's full spectrum of performance.

1.1. Prior work

The W-Score is a quantitative approach and the most prominent metric that uses a "risk-adjusted mortality" provided by the

CONTACT Nicholas J. Napoli njn5fg@virginia.edu Systems and Information Engineering, University of Virginia, 151 Engineer's Way, Charlottesville, VA 22904, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uhse.

TRISS methodology to adjust for patient acuity and evaluate a trauma center's performance (Mock *et al.*, 2004; Osler and Glance, 2007). The W-Score reports an estimate of the number of patients who survived unexpectedly based on their risk of mortality using TRISS. In an attempt to provide a solid systematic and statistical methodology, the W-Score is supported with Flora's Z-statistic and M-statistic (Mock *et al.*, 2004; Boyd *et al.*, 1987; Osler and Glance, 2007). Flora's Z-statistic is a measure of the statistical significance that the W-Score (the number of unexpected survivors), occurred by chance alone (Boyd *et al.*, 1987; Osler and Glance, 2007). The M-statistic measures whether two data sets are similar enough to be compared and produce reliable results. The M-statistic is generated by comparing the proportion of patients in each POS range for both the baseline group and the study group (Boyd *et al.*, 1987). This measure ranges from 0 to 1; as M approaches 1 the two sets of data are considered a match, and as M approaches 0, there is a disparity (Boyd *et al.*, 1987).

1.1.1. Challenges

A trauma center's ideology is designed to provide care for a full spectrum of critical injuries. However, there are several downfalls with the W-Score (including how it is statistically supported), and the perceived interpretation of the metric (Osler and Glance, 2007; Jurkovich and Mock, 1999). Utilizing the W-score produces metric outputs favoring populations dominated by patients with low acuity (>95% probability of survival (POS)) and fails to capture the full spectrum of care provided. The W-score output is highly affected by the amount of people in the study, which is known as the effect size (Jurkovich and Mock, 1999). The W-score becomes inflated as the value of N increases because the distribution of acuity is positively skewed by the relatively larger number of patients with low acuity (Jurkovich and Mock, 1999). Therefore, the number of deaths is small compared to the number of patients admitted to a trauma center, which superficially increases the W-Score (Jurkovich and Mock, 1999). This further supports the problems with using our current metrics, and suggests an alternative metric that is more appropriately fitted for the innate nature of how patient acuity is distributed.

Different levels (POS ranges) of a trauma center's acuity population can dramatically change performance. The distribution of patients' survival rates will vary from center to center. This ultimately limits the usefulness of TRISS when comparing trauma centers that differ in patient acuity levels, demographics, and several other variable factors (Demetriades *et al.*, 2001). Furthermore, these distributions are typically analyzed with small sample numbers to compare yearly or monthly performance changes. Although not thoroughly discussed in the literature, Flora's Z-statistic approach is essentially evaluating the sum of independent binomial random variables with different probabilities. While Flora's Z-statistic is accepted because of its simplicity, this approximation using the normal distribution is not ideal, and other approaches have demonstrated more consistent performance (Butler and Stephens, 1993). Moreover, these approximations for data sets are typically preformed when the sets of probabilities are closely arranged together (Butler and Stephens, 1993; Feller, 1968), which is not the case for Flora's Z-Statistic, where the full range of probabilities are calculated together. When evaluating the entire data set (with the full range of POS values (.0–1.0)) applying Flora's Z-statistic, we encounter the potential for the low–acuity population of patients to attenuate and downplay the outcomes that are associated with the severely acute patients. Essentially, the majority of the probability mass for defining significance is associated with the low–acuity patient population. Therefore, evaluating the center's performance highly favors the low–acuity group and neglects the important patterns among the more critically injured patients. This limits our view of the full spectrum of trauma outcomes.

Furthermore, the M-statistic, which supports the W-Score, is a very weak measure of comparison between the study and baseline patient data sets, and this methodology has not been thoroughly evaluated; thus, its statistical properties and reliability are unknown (Osler and Glance, 2007). Eighty to ninety percent of the patient population in a typical trauma center is comprised of patients with low acuity (>95% probability of survival (POS)). The M-statistic will often report a match, since this percentage will remain consistent among almost all trauma centers. However, this overshadows important differences within the critically injured populations, which are crucial to examine when evaluating a trauma center's performance. The literature reveals numerous correction factors intended to adjust TRISS for more effective predictability (Schluter, 2011; Norouzi *et al.*, 2013; Llullaku *et al.*, 2009; Costa and Scarpelini, 2012; Rogers *et al.*, 2012; Gabbe *et al.*, 2004). The existence of these adjustments further supports the notion that no two trauma centers are the same (Hollis *et al.*, 1995; Demetriades *et al.*, 2001). Moreover, these patient population demographics incorporate variations in the percentage of non-lethal injuries and locality-dependent injury patterns.

This work raises four interesting questions: (1) Is the data truly biased toward the lower acuity patient populations, and how should we think about approaching this issue? (2) What can be done if the current methodology cannot statistically capture changes within crucial patient acuity populations (sub-groups), and can we identify which subgroups require remediation or which have improved? (3) If a metric is supposed to evaluate trauma center performance from the severely critical to the very low-acuity patients, do the current metrics capture the true goals for a trauma center? (4) As metrics are compared over various cohorts, how reliable are these metrics over different sample sizes?

1.1.2. Insights

Despite the limitations of TRISS with its current metrics, TRISS has become a standard utility for objectively evaluating the quality performance of a center (Schluter, 2011; Demetriades *et al.*, 2001; Costa and Scarpelini, 2012). TRISS's strength is in its ability to approximate a patient's systemic acuity level using POS. However, the TRISS metric is concerning for its failure to capture changes in the severely acute patient populations. The concept of the M-Statistic's approach, which stratifies patient sub-groups (acuity levels of different POS ranges), should be leveraged. Each sub-group should be evaluated for their observed and anticipated risk–adjusted mortality in order to develop a more comprehensive performance measure for the full range of POS values. Whether we evaluate mortality based on patient acuity, time period, or with institutional and national benchmarks, the sub-groups of patients under investigation ("acuity sub-groups") must be of sufficient size and integrity to ensure appropriate statistical significance, resulting in relevant conclusions. Therefore, an initial metric would be a performance trend metric that would identify which acuity sub-group in the system requires future improvement as well as which sub-group would provide the most impact to improve performance. In addition, a secondary metric would be normalized by population (where the number of patients within a sub-group does not affect the metric output) to depict a more accurate measure of performance.

1.1.3. Contributions

In this article, we use TRISS to categorize an academic Level 1 trauma center's patient population into an acuity sub-group which captures a statistically significant subset of patients with comparable anticipated survival rates. These survival rates are defined by TRISS's POS, which is a logistic regression that captures multiple physiological variables and categorizes subsets of patients with comparable anticipated survival rates based on this POS. The minimum overall sample POS range required to determine a statistically significant sample is dependent on the number of patients categorized into each POS subset. This work focuses on analytically deriving a new tool that utilizes the same attributes to categorize acuity that TRISS does. We develop a new metric that simultaneously addresses limitations in TRISS and observed mortality while building upon the data created from these current metrics in order to retrospectively evaluate system changes in addition to identifying areas for improvement.

The contributions of this work are:

1. We develop a statistical framework that captures changes within acuity sub-groups.
2. We demonstrate, by analyzing different cohort time periods, that we can identify which specific patient acuity sub-groups have improved as well as which require remediation.
3. We demonstrate how the W-Score can misrepresent a performance index and show that the relative mortality metric (RMM) can evaluate the entire spectrum of care and is unbiased by overrepresented patient acuity sub-groups.
4. We demonstrate that our RMM metric maintains a more enhanced consistency of reliability than the W-Score as sample size changes.

2. Methods

2.1. Data

The collected data is from the trauma registry of a Level 1 trauma center in the United States. A single individual has maintained the registry database at the center over a 20–year period (1994–2013), encompassing 34 735 patient encounters. Although no two trauma centers are the same, the majority of trauma centers are dominated by low–acuity populations similar to our database. The most critical demographic factors that make up a trauma center population include age, injury severity score, and other patient characteristics, which are all accounted for when calculating the POS using TRISS and are shown in Table 1. Table 1 demonstrates 95% confidence intervals (CI). The utilization of a single center's data is a limitation. However, the proposed framework is still applicable for any type of dominant acuity (i.e., low, mild, or severe) and can be extended to multiple trauma centers.

Table 1. Population demographics 1994–2013.

Variable	Total Data Set
Mean Age, $CI_{\alpha=.05}$	42.1 (41.9–42.4)
Percent Female	36.7
Percent Penetrating Injury	8.6
Percent Blunt Injury	91.4
Mean Glasgow Coma Scale, $CI_{\alpha=.05}$	13.6 (9.8–10.4)
Mean ISS, $CI_{\alpha=.05}$	10.1 (9.8–10.4)

Previous work by Napoli *et al.* (2017) explores the potential bias in health records, as well as methods to overcome this bias. This method examines if the missing physiological and anatomical variables required to calculate POS accounting were truly random processes in order to detect bias. The years 2000, 2001, and 2002 demonstrated bias within the records and were deemed unrecoverable. Thus, these years were discarded from the data set and any additional patient encounters in which the POS was unable to be calculated reduced the data set down to 25 575 patients. The proposed method in this work establishes a gap analysis for an ICU which provides a risk adjusted mortality metric and identifies inconsistent patterns by retrospectively examining patient's anticipated mortality over a 20–year period.

2.2. Design overview

This design method first addresses how to properly define patient sub-groups and performance trajectories and then introduces new metrics for evaluation and comparison. A step-by-step framework is discussed:

1. Stratify Into Acuity Sub-groups Using TRISS;
2. Calculating Minimum Acuity Sub-group Size;
3. Defining Relative Mortality Performance Trend (RMPT);
4. Defining Relative Mortality Metric (RMM).

2.2.1. Stratify into patient sub-groups using TRISS

The number of non-survivors within a patient population is meaningless without due consideration to injury severity and anticipated mortality. TRISS provides a repeatable, well-studied measure of anticipated mortality based on both injuries and physiological vitals at patient presentation. This adjusted risk of mortality can be thought as a patient's acuity level and therefore a measurement of the intensity of care required. We quantify the adjusted risk mortality via logistic regression (Eq. (1)) defined as,

$$POS = \frac{1}{1 + e^{b_0 + b_1(RTS) + b_2(ISS) + b_3(AgeIndex)}}, \qquad (1)$$

where variables such us the Revised Trauma Score (RTS), Injury Severity Score (ISS), and Age Range (Age Index) are accounted for within the model. Utilizing TRISS POS ranges, both populations (survivor and non-survivor) are sorted into similar acuity sub-groups (bins). Thus, we calculate an observed mortality, O_b, by,

$$O_b = \frac{D_b}{N_b}, \tag{2}$$

where N_b is the total number of patients contained in bin, b, and D_b is the number of non-survivors in bin, b. When repeated over the entire range of acuity sub-groups, this provides an overview of non-survival trends by patient acuity. In the following, we address calculations of minimum statistically significant bin size; a concern due to the finite number of trauma patients.

2.2.2. Calculating minimum sub-group size

Confidently reporting and validating the strength of each O_b requires careful sizing of its specific b. We are met with a trade-off where large bin ranges will cause a reduction in resolution of acuity sub-groups, but provide an increase in statistical confidence. On the other hand, smaller bin ranges would provide increased resolution, but with the cost of losing statistical confidence for that bin's O_b.

In order to enhance resolution while maintaining a defined statistical confidence, we adjust bin intervals dynamically based on POS population distribution. The goal of this adjustment is to overcome the inherent limitations of applying a linear binning model to a population distribution of POS values that does not fit a uniform distribution for which a linear binning would be appropriate. We represent each bin interval by the POS value at its geometric midpoint. This is desirable for both simplicity of calculation and inter-institutional comparisons. Table 2 indicates the minimum bin size required, regardless of the the probability value, to maintain error below the given threshold (see later in the discussion for particular caveats when using the binomial approximation for when probabilities of survival are extremely low or high). Rather than using arbitrary bin sizes, patient acuity resolution can be maximized by defining the minimum bin sizes necessary to obtain sample sizes appropriate for the desired confidence interval.

For each b, we must quantify the minimum number of samples, N_b, required. The bin interval provides the range of POS values within b. Using this information and the midpoint of the bin interval, we can determine the anticipated probability of mortality within b as $1 - POS_b = A_b$. For example, if we examine the bin associated with a POS of 20% we would anticipate 80% of the patients within that bin not to survive. Using this probability, we can determine an adequate sample size similar to the way we would calculate the number of trials required to statistically determine if a coin is fair. Therefore, a binomial distribution is used and then approximated to a normal distribution. Using the normal approximation, the standard error is defined as

Table 2. Minimum bin size.

Error	Bin Size
6%	≥ 267
8%	≥ 150
10%	≥ 101
12%	≥ 101

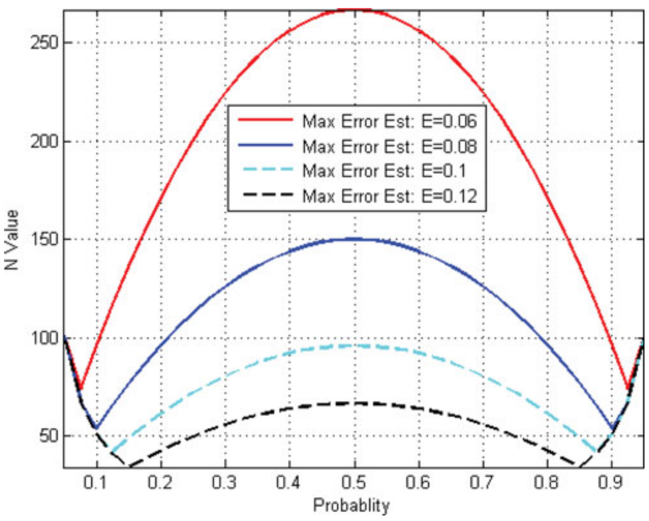


Figure 1. N Values with Z = .95. This curve depicts the minimum N value for specific confidence levels as probability of survival changes.

$$S_b = \sqrt{\frac{A_b(1 - A_b)}{N_b}}, \tag{3}$$

Using the standard error, we define the maximum error (confidence interval bound), E_b, by

$$E_b = Z\sqrt{\frac{A_b(1 - A_b)}{N_b}}, \tag{4}$$

where Z sets the confidence level. This would allow us to inversely calculate the appropriate required N value for that specific bin by,

$$N_b = \frac{Z^2(A_b(1 - A_b))}{E_b^2}. \tag{5}$$

Due to the nonlinear effect A_b has on N, depending on the confidence level we impose, N can change drastically across various bins shown in Fig. 1 and Table 3. In addition to this minimum constraint on N_b from the standard error, there is an additional constraint or rule of thumb on choosing N when a binomial distribution is approximated. This states that $N \cdot A_b$ and $N(1 - A_b)$ must be greater than 5 (Brown *et al.*, 2001; Moore, 2003). In Fig. 1, we incorporated both criteria to

Table 3. Minimum bin size by POS and error tolerance.

POS	6%	8%	10%	12%
0.05	101	101	101	101
0.15	136	77	49	34
0.25	200	113	72	50
0.35	243	137	87	61
0.45	264	149	95	66
0.50	267	150	96	67
0.65	243	137	87	61
0.75	200	113	72	50
0.85	136	77	49	34
0.95	101	101	101	101

determine the minimum value of N as a function of A_b. Once a minimum confidence or maximum error is decided, the peak value corresponding to that error curve becomes our threshold value for the nonlinear binning process which follows.

2.2.3. Defining relative mortality metric

The relative mortality metric (RMM) describes the overall performance in relation to the anticipated mortality (benchmark TRISS threshold from the national database) and observed mortality, where RMM has a range from −1 to 1. When the RMM is zero, the expected TRISS national benchmark outcome is achieved for the entire acuity spectrum. If the RMM value is greater than zero, the center is outperforming the anticipated national benchmark set by TRISS. Likewise, when the RMM is negative, we are underperforming as compared to the benchmark. Therefore, the RMM is to provide a quantitative metric of how an institution's overall performance has adjusted, providing a single metric for performance and its confidence intervals. RMM is defined by

$$RMM = \frac{\sum_{b=1}^{j} R_b(A_b - O_b)}{\sum_{b=1}^{j} R_b(A_b)} \tag{6}$$

$$= 1 - \frac{\sum_{b=1}^{j} R_b O_b}{\sum_{b=1}^{j} R_b(A_b)} \tag{7}$$

where R_b is the bin range interval associated with b, the bin index, A is the anticipated probability of mortality, and O is the observed probability of mortality. The lower and upper limits of the RMM (RMM_{UL} and RMM_{LL}, respectively) are noted by

$$RMM_{UL} = 1 - \frac{\sum_{b=1}^{j} R_b(O_b + E_b)}{\sum_{b=1}^{j} R_b(A_b)}, \tag{8}$$

$$RMM_{LL} = 1 - \frac{\sum_{b=1}^{j} R_b(O_b - E_b)}{\sum_{b=1}^{j} R_b(A_b)} \tag{9}$$

where the confidence level for each bin, E_b, is applied to determine significance between different RMM scores.

2.2.4. Relative mortality performance trend

The Relative Mortality Performance Trend (RMPT) extends the Relative Mortality Analysis (RMA) by providing a picture of trends over time rather than being limited to a single value metric and by providing insight into specific patient populations (acuity sub-groups). This allows two comparative gap analyses: (1) the RMPT allows us to examine how an institution changes with respect to itself in terms of time, patient conditions, treatment, etc., so we can identify opportunities for improvement within acuity sub-groups that may have been overlooked. (2) The RMPT allows us to target specific acuity sub-groups to evaluate performance at a set benchmark (i.e., anticipated mortality). While TRISS anticipated mortality (RMM=0) is a convenient reference point, there are obvious discrepancies that occur with a center's ability to truly match the national database across all aspects (i.e., demographics, types of injuries). Thus, it should not be the primary point of comparison. Instead, the RMPT should be used as a tool to determine center performance goals based on their own trends at various acuity sub-groups. The RMPT is developed by plotting each individual observed mortality bin and their respective confidence interval for each cohort.

3. Results and discussion

In this section, we evaluate and validate the performance of the proposed Relative Mortality Analysis methodology for evaluating trauma center performance using the following four research questions:

RQ1: Are there overrepresented patient groups, and does it matter how we stratify these patients into sub-groups?
RQ2: Does our statistical framework capture changes within patient acuity sub-groups over different cohorts of time in order to identify changes in performance?
RQ3: Does the W-Score misrepresent an institution's index of performance and how does the RMM compare?
RQ4: What is the reliability of the RMM and W-Score metrics when the sample size isn't large enough to represent the actual patient population?

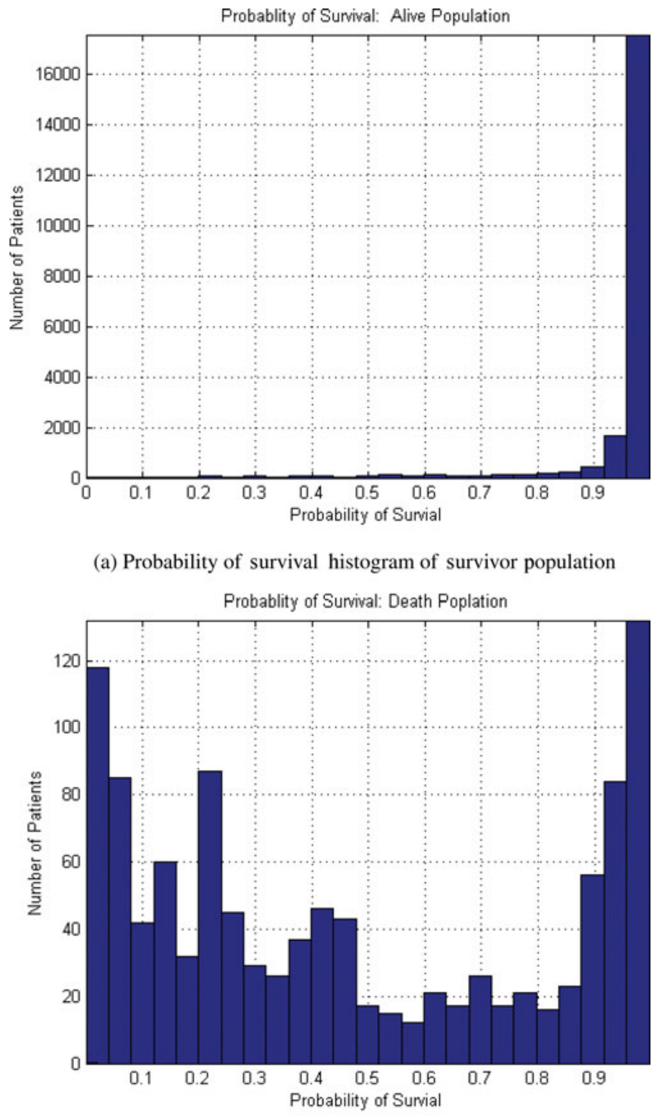
3.1. RQ1: Patient acuity sub-groups

We address this research question in two parts by: (1) examining the POS distribution of patients; and (2) discussing the statistical impact of stratifying these patient acuity sub-groups. Although some literature has pointed to overrepresentation of low–acuity patients in trauma data sets, this is an integral discussion point and the cornerstone for this work. This is a paramount issue because there is currently no statistical methodology to address this problem.

A proper representative metric should support the center's performance across the entire spectrum of injury severity without prejudice toward an overrepresented patient acuity sub-group, such as patients with high POS. Analysis of overrepresented acuity groups where the number of people are incorporated into the direct metric output places a heavier weight towards that specific acuity group, reducing the ability to capture other acuity performance changes within the population. As such, our work on defining a patient sub-group and understanding the trade-offs of stratifying these sub-groups is paramount.

3.1.1. Are there overrepresented patient acuity sub-groups?

The full data set was split based on patient outcome (non-survivor and survivor), and the POS for each patient was calculated using Eq. (1) to provide insight into their acuity level. Histograms of patients' POS were generated for both the survivor and non-survivor populations, which can be seen in Figs. 2a and 2b. We can quickly note, by examining the non-survivor and survivor histograms in Figs. 2a and 2b, that these distributions are not uniform. The lower acuity patient sub-groups ($POS > .9$) have an overwhelming presence within the 23 454 patient records, which comprises 89.39% of the data. These findings are analogous to the reported patient POS distributions in the trauma literature, demonstrating overwhelming overrepresentation of patients with high probability of survival (low acuity) (Boyd *et al.*, 1987; Hollis *et al.*, 1995; Bouillon *et al.*, 1997; Demetriades *et al.*, 2001; Hariharan *et al.*, 2009). Thus, approximately only 10% of the patient population captures



(a) Probability of survival histogram of survivor population

(b) Probability of survival histogram of non-Survivor population

Figure 2. The distributions of survivor and non-survivor populations as a function of their calculated probability of survival (POS).

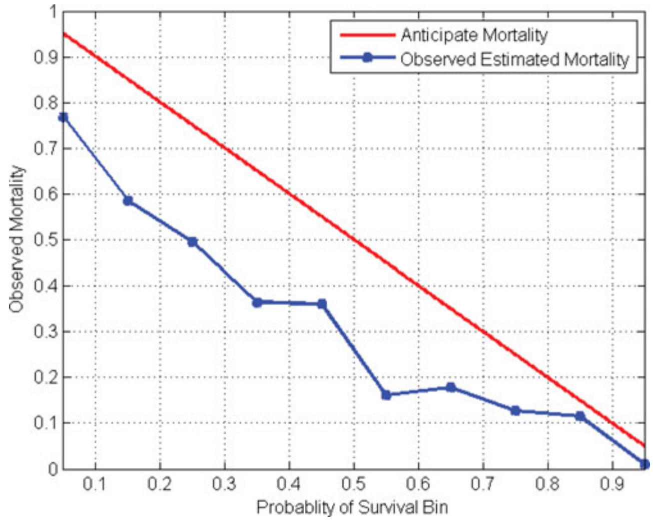


Figure 3. Normalization of probability of survival with linear binning.

approximately 90% of the actual full spectrum of care (POS from 0 to .9). Hence, we can see that, when running the standard statistical methodology, we are not evaluating the true performance of a center, but how a center performs with the low–acuity patients, a sub-group of the actual population. Therefore, this innate disparity between the severely critical impact on metric performance is attenuated and must be addressed.

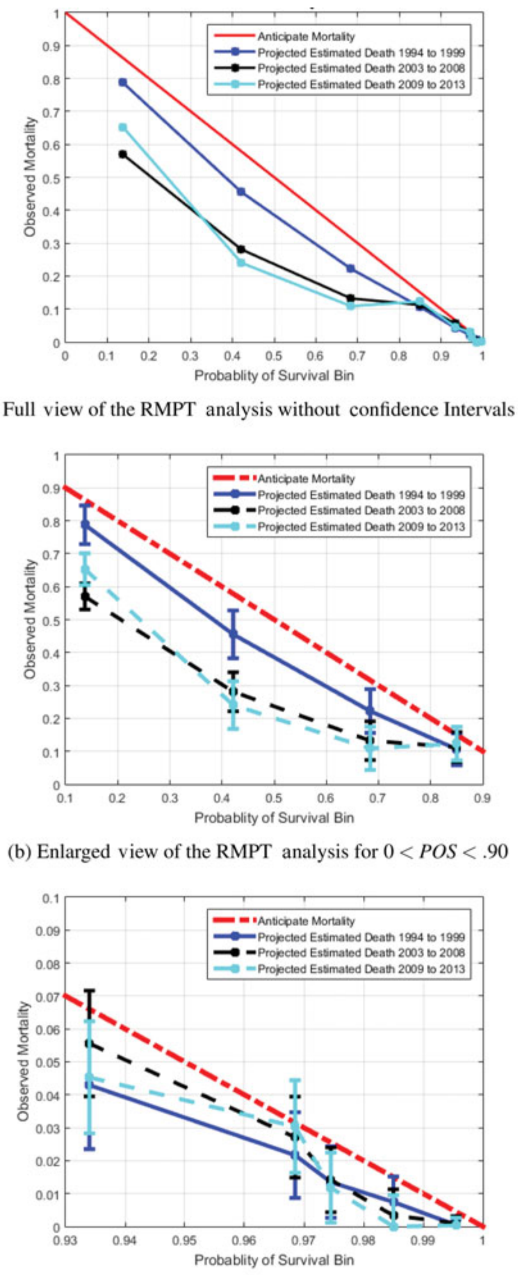
3.1.2. Stratifying patient acuity sub-groups and understanding the trade-offs

In Fig. 2b, we note an alarming increase in the non-survivors in the patient sub-group for the bins associated with high POS. Thus, examining total mortality provides poor insight and is an inappropriate measure of performance. However, normalizing the number of deaths by the total number of patients associated within a sub-group would provided a meaningful value. This value is defined as the observed mortality described in Subsection 2.2.1 and seen in Fig. 3. Additionally, the red line shown in Fig. 3 is the anticipated mortality, which was the benchmark associated with patient sub-groups based upon TRISS. Utilizing Fig. 3, we are then able to evaluate these acuity sub-groups and gain insight into their anticipated versus observed mortality for a given patient acuity.

However, the usefulness of these observations is limited without knowing the confidence of those observations, as described in Section 2.2.2. Thus, as the sample size increases, we can gain statistical power and better refine our confidence intervals, but in order to gain a higher sample size, the POS range must be increased for a defined acuity sub-group. Therefore, we are presented with a trade-off between having a more refined confidence interval and having a more granular patient acuity sub-group.

When this overrepresentation in an acuity sub-group is present, it is not possible to specify a single defined bin width that would provide adequate statistical power for low POS acuity sub-groups and reasonable resolution of acuity sub-groups with high POS values across the entire spectrum of acuity. The nonlinear bin width method has the ability to further increase the granularity of acuity sub-groups with like acuity and maintain statistical power. This is demonstrated by examining the two different binning methods present (linear and nonlinear) in Figs. 3 and 4a. Moreover, in Tables 4 and 5, we can note how a single acuity sub-group is represented by 89.39% of the data set versus 48.68%. The bias of acuity must be recognized and dealt with accordingly, which the current literature does not address (Hollis *et al.*, 1995; Joosse, 2013).

However, acuity sub-groups of like acuity should be more formally addressed and understood. Though POS can be thought of as a continuous variable, for the purposes of analysis, the values are discretized to the thousandth decimal within our data set of 24 000 patients. When examining large numbers of patients within a small range of POS values, we are provided with a plethora of patients clustered together with identical POS values. This explains why the second partitioned group of quartiles ($POS > .9$) is not perfectly broken into quartile groups, seen in Table 5. Addressing the various permutations of TRISS attributes a patient can obtain to achieve an identical POS value should be conceptualized. For improvement and understanding of these acuity sub-groups, we must inquire how



(a) Full view of the RMPT analysis without confidence Intervals (CI)

(b) Enlarged view of the RMPT analysis for $0 < POS < .90$

(c) Enlarged view of the RMPT analysis for $POS > 0.9$

Figure 4. Overview of the RMPT analysis.

truly distinct is a POS acuity sub-group of exactly .997, and what is the clinical significance to this distinction?

Minimum patient sub-groups were calculated as discussed in Section 2.2.2 using nonlinear binning. In the histogram of trauma patients population shown in Fig. 2a and in Table 4, one can note that there is a discontinuity at the probability of survival $\geq$ 90%. We choose to treat it as a piece-wise function, where $[0 - .9)$ demonstrates a closeness to a uniform behavior and $[.9 - 1]$ demonstrates an exponential behavior. This discontinuity in the histogram is used to build the nonlinear binning framework. Therefore, sub-group–based quartiles were introduced to each domain of the piece-wise function.

Furthermore, analysis indicated a sizable population with a POS $\geq$ 99.9%, where we could have extended more bins into the population for enhanced resolution of acuity. However, we should note, when examining bins at the extremes of the data where the probability is close to zero or one, the approximations work poorly and a larger N value is needed (Moore, 2003). Therefore, only a single bin was created, bin 9, to adequately represent the sample population for this probability close to one. Compared to previous methods (Hollis *et al.*, 1995), this provides both greater resolution within the high POS portion of the range and more explicit accuracy. These bins are asymmetrically scaled, with bins narrowing as they approach a POS of 1 for improved resolution, as shown in Table 5.

Table 4. N values linear bins of complete data set.

Sub-Group		Outcome N Value			
Bin	POS Range	S	D	N Value	Percent of % of set
1	0.0–0.1	59	196	254	1.08%
2	0.1–0.2	68	96	165	0.70%
3	0.1–0.2	138	135	273	1.16%
4	0.3–0.4	111	63	174	0.74%
5	0.4–0.5	147	83	230	0.98%
6	0.5–0.6	166	32	198	0.84%
7	0.6–0.7	242	52	294	1.25%
8	0.7–0.8	246	36	282	1.20%
9	0.8–0.9	547	71	618	2.63%
10	0.9–1.0	20745	221	20966	89.39%

We can note, through our data and other published work, that it is typical to receive distribution of acuity that favors the low–acuity population. The use of stratifying the patient population in a nonlinear fashion allows for the examination of smaller clusters of the patient populations (acuity sub-groups) and maintains a larger sample size for a stronger statistical confidence of their observed mortality.

3.2. RQ2: Analyzing different cohort time periods using RMPT

The ability to depict a proper trajectory over a cohort of years for ICU performance provides a means to examine patterns of performance trends that compares like-acuity sub-groups over time and to a set benchmark (TRISS's anticipated mortality). Using the established confidence levels for each observed mortality allows us to examine an institution's performance on numerous sub-groups of like-acuity patients over time.

It is worth noting that, in our analysis for which we compared different cohort years, we utilized the geometric midpoint of the bin as the observed mortality of the entire bin for computational simplicity and symmetry. When binning in this manner, it is

Table 5. Nonlinear bins: N-values decomposed.

Sub-Group		Outcome N Value			Cohort N Value			
Bin	POS Range	S	D	Total N	94–99	03–08	09–13	Percent of the Set
1	0.000–0.274	222	399	621	132	293	196	2.65%
2	0.274–0.569	425	199	624	178	267	179	2.66%
3	0.559–0.799	529	95	624	189	242	193	2.66%
4	0.799–0.901	554	71	625	207	231	187	2.66%
5	0.901–0.967	2229	115	2364	629	918	817	10.08%
6	0.967–0.970	2005	54	2059	692	774	593	8.78%
7	0.970–0.979	2617	35	2652	816	988	848	11.31%
8	0.979–0.991	2457	10	2467	942	899	626	10.52%
9	0.991–1.000	11411	7	11418	4364	3884	3170	48.68%

Table 6. Evaluation of bins for midpoint bias.

Bin	Geometric Midpoint	Mean '94–'99	Mean '03–'08	Mean '09–'13
1	0.1370	0.1645	0.1595	0.1615
2	0.4215	0.4320	0.4245	0.4376
3	0.6840	0.6918	0.6914	0.6959
4	0.8500	0.8595	0.8551	0.8598
5	0.9340	0.9433	0.9444	0.9436
6	0.9685	0.9680	0.9680	0.9680
7	0.9745	0.9760	0.9761	0.9761
8	0.9850	0.9859	0.9861	0.9862
9	0.9955	0.9947	0.9947	0.9947

adequate to simplify by accepting the geometric midpoint of the bin as the bin value. However, the confidence intervals of these bins are not precise, since the observations within the bin are only a representative sample. Therefore, this approach introduces a potential skew in the resulting trend, which is prudent to examine before accepting this simplification, shown in Table 6. This simplification is beneficial predominantly for the kind of temporal comparison (different cohort years) allowing for direct comparison between cohorts and acuity sub-groups. However, if there are concerns about using the geometric midpoint in order to define the confidence intervals, another approximation for defining the maximum error (confidence interval bound), E_b, is done by

$$E_b = Z\sqrt{\frac{\sum_{i=1}^{n} O_{ib}(1 - O_{ib})}{N_b}}, \tag{10}$$

where Z sets the confidence level, b is the bin, and i is the index for the patient associated with that bin. Although this approximation it a bit more computationally complex, it will better approximate the confidence interval when we have a skewed geometric mean within a bin. The concern with the midpoint method is the potential of how the observed mortality can occur in minor shifts due to the use of the geometric mean instead of the population mean for a given mortality bin, shown in Table 6, creating a slight acuity imbalance within the trend.

Using the midpoint, sub-groups' Relative Mortality Performance Trends for the three cohort years are presented in Fig. 4a. Figures 4b and 4c provide an enhanced view with confidence intervals demonstrating significant changes in our institute's observed mortality for bins 1 and 2 for (2003–2008) and (2009–2013) when compared against the (1994–1999) cohort. There is no significant improvement between any observed mortality acuity sub-group between the (2003–2008) and (2009–2013) cohort years.

In Figs. 4a and 4c, a strong convergence of the trends towards the anticipated mortality benchmark at bin 4 (midpoint = .85) for all the cohorts is observed. We note this as a pivotal shift in the trend, since bins 1, 2, and 3 do not incorporate the anticipated mortality within their confidence intervals for all the cohort years. Using the benchmark and confidence intervals in Fig. 4c, we can note that for bin 4 (midpoint = .85) and bin 6 (midpoint = .9685), for all the cohort years, they have not been able to demonstrate a significant deviation of a lower mortality from the anticipated mortality. However, bins 7, 8, and 9 have maintained their confidence intervals outside the range of the anticipated mortality, performing better than the benchmark.

Table 7. Observed mortality (O.M.) and standard error (S.E.).

Sub-Group		Years 94–99		Years 03–08		Years 09–13	
Bin	POS Range	O.M.	S.E	O.M.	S.E	O.M.	S.E
1	0.000–0.274	.7879	.0587	.5700	.0394	.6531	.0481
2	0.274–0.569	.4551	.0725	.2809	.0592	.2402	.0723
3	0.559–0.799	.2222	.0663	.1322	.0586	.1088	.0656
4	0.799–0.901	.1063	.0486	.1126	.0460	.1230	.0512
5	0.901–0.967	.0429	.0194	.0556	.0161	.0453	.0171
6	0.967–0.970	.0217	.0130	.0271	.0123	.0304	.0141
7	0.970–0.979	.0135	.0108	.0142	.0098	.0118	.0106
8	0.979–0.991	.0074	.0078	.0033	.0078	0	.0095
9	0.991–1.000	.0005	.0020	.0010	.0021	.0003	.0023

The calculations shown in Table 7 are the observed mortality found in Figs. 4a, 4b, and 4c. In addition, the confidence intervals using Eq. (4) with an assigned Z value for 95% confidence, produced Table 4a and Fig. 4c.

In the past, TRISS has been proved to be a tool for quality analysis in investigating deaths (Costa and Scarpelini, 2012; Karmy-Jones *et al.*, 1992; Masella *et al.*, 2008), which is one of the benefits of parceling acuity sub-groups of TRISS for trajectories. We can see from Table 7 and Fig. 4a these pivotal changes in trajectory. When comparing trends over time, a significant change occurs in the trend for bins 1 and 2 (POS < 0.569), depicting a reduction of mortality for the cohort years of 2003–2013. Although there is no other significant changes between mortality trends over time for other bins, we have been able to outperform the benchmark for bins 1, 2, 3, 7, and 9 for all the cohort years. Likewise, the like-acuity bins of 4 and 6 have neither improved over time, nor have they been able to outperform the set benchmark of anticipated mortality. These acuity sub-groups 4 and 6 are the anomalies within the performance trajectory that have been identified for further investigation, in hopes to explain their stagnation for improvement or simply any change over the last 20 years. Through this analysis of the relative mortality performance trends, it is apparent that we can target, track, and evaluate numerous patient acuity sub-group's performance over time. This type of analysis is key for identifying potential, specific, patient acuity populations that have potential for improvement.

3.3. RQ3: W-Score vs. the relative mortality metric

In the previous research questions, we have demonstrated that there are times when the RMM and W-Score tend to have similar trajectories when compared to different cohorts of time (Table 8). However, this is not always the case. In order to examine if and how the W-Score misrepresents performance and how this compares to the RMM, we examine a subset of the data. The data consists of 1719 patients with an injury severity score (ISS) of 25 or greater in a single trauma center that are associated with two different cohort times (Period A: 1994–1999 and

Table 8. W-Score vs. RMM compared over different time periods.

Metric W-Score	Years 94–00 1.33	Years 03–08 3.01	Years 09–13 2.77
RMM_{LL}	0.066	0.383	.384
RMM	0.173	0.440	.447
RMM_{UL}	0.259	0.487	.498

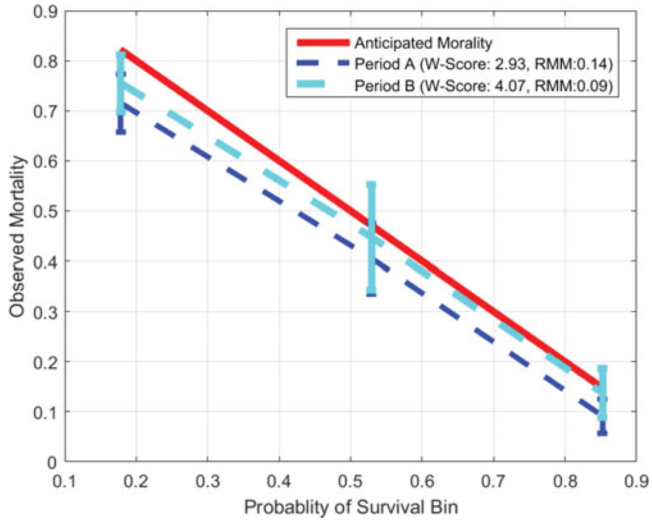


Figure 5. Relative mortality performance trend for ISS 25 or greater, demonstrating that neither cohort is significantly different over the full spectrum of care.

Period B: 2009–2013). We used our relative mortality metric (RMM) and W-Score to conduct performance analysis amongst 1719 patients treated in a university Level 1 trauma center. We compared the performance using the two methodologies: RMM and W-Score methodology for the two cohort time periods.

Using the W-Score, we found a 39.09% improvement in unexpected survivors between the two time periods (Period A: +2.933, Period B: +4.0795). However, the RMM Metric demonstrates the polar opposite; a 55.15% decrease in performance (Period A: 0.1543, Period B: 0.0692) at every level of acuity. Figure 5 depicts the Relative Mortality Performance Trend for Periods A and B. We can see that, during Period B, overall performance is worse, which is accurately reflected in the RMM metric but not in the W-score. While the difference in performance with the Relative Mortality Method is not statistically significant, it still provides a more clear and accurate picture of the direction in overall trends in performance than the W-Score.

Neither metric is truly misrepresenting the index of performance. We must ask ourselves what we really want to measure. The W-Score is a metric that evaluates the entire population of an institution, which only truly describes how an institution's performance is for the entire patient population as a whole. Thus, if our population as a whole does better, the performance will demonstrate an increase. The RMM is a method that combats this issue by accounting for all levels of acuity with equal weight, thus capturing the full spectrum of care. Rather than capturing the performance of a single, dominating acuity population, the RMM allows more accurate discrimination of performance across the entire range acuity. A tool that allows improved discrimination leads to more targeted performance improvement and better utilization of resources. We recommend using this tool to examine longitudinal mortality gap analysis in trauma centers.

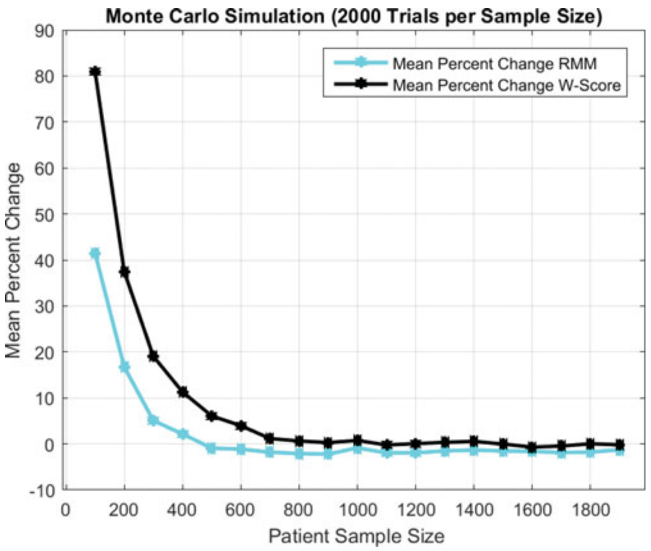


Figure 6. Evaluating the effects sample size has on metric reliability by utilizing Monte Carlo simulations to examine percent change error.

3.4. RQ4: Effects of sample size on RMM and W-Score reliability

When routinely evaluating processes and performances at the clinical level in short time increments, or simply when the volume of patients is small, the sample size may not adequately represent the distribution of patient acuities (POS values). This produces discrepancies in the reported metric between the population score and the score of the data available. As far as we are aware, there is no present literature exploring how the sample size can potentially alter the W-Score metric output. Thus, we generated a Monte Carlo simulation to examine and compare these RMM and W-Score inconsistencies for different sample sizes. Since these metrics are potentially altered as the sample size changes, it is critical to understand how this metric can be skewed.

We first calculate the RMM and W-Score over the entire cleaned data set of 25 757. These two values are considered the actual true performance of a trauma center referred to as RMM_{ref} and $W - Score_{ref}$. We then apply the Monte Carlo simulation by randomly taking a sample size from the full data set over 2000 iterations (sample sizes range from 100 to 2000 patients). The significance of the sample size for each iteration is tested using the criteria published for the W-Score (Peitzman *et al.*, 2012; Flora, 1978), where only the significant iterations were utilized. We use the mean metric scores of the 2000 iterations for each sample size and calculate the percent change between these means and the reference metrics (RMM_{ref} and $W - Score_{ref}$). This allows us to compare both metrics on a similar scale. Figure 6 depicts this mean percent change over each sample size, which gives us an indication of the accuracy of each metric given at different sample sizes. We can see that when the sample is small, using RMM more accurately predicts the true population metric twice as well as the W-score. This approximate 2x factor of improvement can be observed for the sample sizes from 100 to 600 before it begins to converge. This finding can be concerning if institutions use the W-Score when they only have a few hundred patients per year. Therefore, when low sample size conditions are present, RMM is a more reliable reported metric than the W-Score.

Furthermore, the confidence intervals of the RMM will depict this uncertainty in the score measurement and can be reduced by altering the binning scheme. The binning scheme of the two quartile measurements that was presented in the earlier article was used across the entire range of sample sizes. The reported RMM percent change within the simulation could potentially be further improved. This can be accomplished by adjusting the binning for specific sample size and distributions. Therefore, we would potentially provide a higher precision for the metric by reducing the confidence intervals, discussed earlier.

4. Conclusion

In summary, our trauma center, over time, has become increasingly effective in reducing mortality in their targeted population of patients with devastating traumatic injuries. This progressive divergence from the anticipated mortality demonstrates that the implementation of standardized trauma procedures benefits first, and most significantly, the critically injured. Unfortunately, these strides in caring for the most critical patients have not been matched by progress in caring for the more common patients: those with $POS > 0.569$. The RMM and W-Score both have demonstrated identical trends of initial improvement. However, the RMM has proved robust in analyzing biased data and evaluating all levels of acuity using the Relative Mortality Performance Trends to better summarize performance for the full spectrum of care. Future research will focus on intervals depicting irregularities within the existing trends over time, during which there was no improvement among specific patient acuity sub-group bins 4 and 6. These acuity sub-groups are compelling both in the volume of patients and unexpected mortality rates. This establishes a potential for crucial improvement, given that the causal agents are controllable.

The validation of RMM demonstrates strong functionality in terms of evaluating performance across the entire domain of acuity, rather than being influenced by bias from a specific acuity sub-group. The W-Score is a problematic metric in which there is no solid framework that addresses the distribution of the POS values, or the N values. RMM fully addresses the ICU in a systemic means, where every acuity sub-group is treated equally, regardless of the amount of patients within the bin. Although a random incoming patient will more likely be associated with one particular bin because of how the POS is distributed, we provide a means of evaluating the intensive care unit as a whole for all levels of acuity. Therefore, it is obvious, with the inverse exponential distribution of POS values to the right (Fig. 2a), that the W-Score struggles with the ability to definitively measure a trauma center's ability to effectively save patients who are severely injured. Compared to the W-Score, the RMM method provides a notably greater emphasis on care provided to the high–acuity (low POS) patient population as well as greater resolution within the low–acuity (high POS) patient population. Additionally, Fig. 6 demonstrates how we can achieve a decreased error in our reported performance when using RMM with a low sample size, providing a more accurate and robust metric.

The W-Score provides an overall view of trauma center performance by considering each patient as partially dead and partially alive (a fractioned value of a person); then, it compares the summation of predicted living and dead patients to the observed numbers. Although it is a convenient computational method, it provides neither the confidence that the overall result is statistically significant from previous years, nor the ability to statistically examine performance within acuity sub-groups. This is particularly concerning when we have a low sample size for a center with high–acuity patients (low POS values). This prevents us from converging to an actual sample mean (observed mortality) for high–acuity patients, and the W-Score weights them inappropriately as a comprehensive metric. In order to examine performance within a subset of patients (i.e., those severely injured), that subset must contain enough patients to minimize chance variation. The method described in Section 2.2.2 ensures this, while the W-Score does not.

Prior trauma center benchmarks that have used the W-Score are heavily influenced by the disproportionately low cases of patients with high acuity, discrepancy from national norms, and inconsistencies in their statistical rigor (Hollis *et al.*, 1995; Hariharan *et al.*, 2009; Joosse, 2013). This work presents a method for processing TRISS scores to yield a statistically significant and high–resolution picture of the evolution of an institution's care over time.

Advantages of this method include avoiding demographic discrepancy, overcoming non-uniform distributions of POS values where specific acuity sub-groups have an overwhelming representation, incorporating statistical confidence, proper development of temporal trends, and a comprehensive performance metric to view the entire spectrum of trauma care. Overall, the insight provided by this method will lead to more accurate depictions of trauma center performance and, ultimately, a better understanding of opportunities for future center improvement.

Funding

A special thanks to the Walter and Martha Curt Foundation for their support of this research.

ORCID

Nicholas J. Napoli http://orcid.org/0000-0002-9071-3965

References

Biffl, W. L., Harrington, D. T., Majercik, S. D., Starring, J., and Cioffi, W. G. (2005) The evolution of trauma care at a Level 1 trauma center. *American College of Surgeons*, **200**, 922–929.

Bouillon, B., Lefering, R., Vorweg, M., Tiling, T., Neugebauer, E., and Troidl, H. (1997) Trauma score systems: Cologne validation study. *J Trauma*, **42**, 652–8.

Bouzat, P., Ageron, F., Brun, J., Levrat, A., Berthet, M., Rancurel, E., Thouret, J., Thony, F., Arvieux, C., Payen, J., and Group, T. (2015) A regional trauma system to optimize the pre-hospital triage of trauma patients. *Critical Care*, **19**, 1–9.

Boyd, C., Tolson, M., and Copes, W. (1987) Evaluating trauma care: The triss method. *J Trauma*, **27**, 370–8.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001) Interval estimation for a binomial proportion. *Statistics Sci.*, **16**, 101–133.

Bulter, K., and Stephens, M. (1993) *The Distribution of Sum of Binomial Random Variables*. Office of Naval Research: Technical Report 467. Department of Statistics, Stanford University: Stanford, CA.

Costa, C., and Scarpelini, S. (2012) Evaluation of the quality of trauma care service through the study of deaths in a tertiary hospital. *Rev Col Bras Cir*, **39**, 249–54.

Demetriades, D., Chan, L., Velmanos, G., Sava, J., Preston, C., Gruzinski, G., and Berne, T. (2001) Triss methodology: An inappropriate tool for comparing outcomes between trauma centers. *J Am Coll Surg*, **193**, 250–4.

Feller, W. (1968). The Binomial and the Poisson Distributions, in *An Introduction to Probability Theory and Its Applications*. Chapter 6. John Wiley: New York.

Flora, J. D. (1978) A method for comparing survival of burn patients to a standard survival curve. *The Journal of Trauma*, **18**, 701–705.

Gabbe, B., Cameron, P., and Wolfe, R. (2004) Triss: Does it get better than this?. *Acad Emerg Med*, **11**, 181–6.

Hariharan, S., Chen, D., Parker, K., Figari, A., Lessey, G., Absolom, D., James, S., Fraser, O., and Letsholathebe, C. (2009) Evaluation of trauma care applying triss methodology in a caribbean developing country. *J Emerg Med*, **37**, 85–90.

Hollis, S., Yates, D., Woodford, M., and Foster, P. (1995) Standardized comparison of performance indicators in trauma: A new approach to case-mix variation. *J Trauma*, **38**, 763–6.

Joosse, P. (2013) An evolution of trauma care evaluation: A thesis on trauma registry and outcome prediction models, Phd thesis, University of Amsterdam, Amsterdam, The Netherlands.

Jurkovich, G. J., and Mock, C. (1999) Systematic review of trauma system effectiveness bases on registry comparsions. *Journal of Trauma*, **47**, S46–S55.

Karmy-Jones, R., Copes, W., Champion, H., Weigelt, J., Shackford, S., Lawnick, M., Rozycki, G., Hollingsworth-Fridlund, P., and Klein, J. (1992) Results of a multi-institutional outcome assessment: Results of a structured peer review of triss designated unexpected outcomes. *The Journal of Trauma: Injury, Infection, and Critical Care*, **32**, 196–203.

Llullaku, S. S., Hyseni, N. S., Bytyci, C. I., and Rexhepi, S. K. (2009) Evaluation of trauma care using triss method: The role of adjusted misclassification rate and adjusted W-statistic. *World Journal of Emergency Surgery*, **4**, 1–6.

Masella, C. A., Pinho, V. F., Passos, A. D. C., Netto, F. A. C. S., Rizoli, S., and Scarpelini, S. (2008) Temporal distribution of trauma deaths: Quality of trauma care in a developing country. *The Journal of Trauma: Injury, Infection, and Critical Care*, **65**, 653–766.

Mock, C., Lormand, J., Goosen, J., Joshipura, M., and Peden, M. (2004) Guidelines for essential trauma care. World Health Organization: Geneva. http://apps.who.int/iris/bitstream/10665/42565/1/9241546409_eng.pdf.

Moore, D. S. (2003) *The Basic Practice of Statistics with CDROM*. W. H. Freeman & Co.: New York, NY.

Napoli, N., Kotoriy, M., Barnhardt, W., Young, J., and Barnes, L.E. (2017) Addressing bias from non-random missing attributes in health data. *IEEE International Conference on Biomedical and Health Informatics*, 1–4.

Norouzi, V., Feizi, I., Vatankhah, S., and Pourshaikhian, M. (2013) Calculation of the probability of survival for trauma patients based on trauma score and the injury severity score model in fatemi hospital in ardabil. *Arch Trauma Res*, **2**, 30–5.

Osler, T., and Glance, L. G. (2007) Evaluating Trauma Center Performance. In Flint, L., Meredith, J. W., Schwab, W., Trunkey, D. D., Rue, L., and Taheri, P. A., Eds. *Trauma Contemporary Principles and Therapy*, Chapter 16, pp. 191–199. Lippincott: Philadelphia.

Peitzman, A. B., Courcoulas, A. P., Stinson, C., Udekwu, A. O., Billiar, T. R., and Harbrecht, B. G. (2012) Trauma center maturation quantification of process and outcome. *Annals of Surgery*, **230**, 87.

Petrie, D., Lane, P., and Stewart, T. (1996) An evaluation of patient outcomes comparing trauma team activated versus trauma team not activated using triss analysis. *J Trauma*, **41**, 870–3.

Rogers, F., Osler, T., Krasne, M., Rogers, A., Bradburn, E., Lee, J., Wu, D., McWilliams, N., and Horst, M. (2012) Has TRISS become an anachronism? A comparison of mortality between the national trauma data bank and major trauma outcome study databases. *J Trauma Acute Care Surg.*, **73**, 326–31.

Schluter, P. (2011) The trauma and injury severity score (TRISS) revised. *Injury*, **42**, 90–6.