SPECIAL ARTICLE

Methodology and Analytic Rationale for the American College of Surgeons Trauma Quality Improvement Program

Craig D Newgard, MD, MPH, FACEP, John J Fildes, MD, FACS, LieLing Wu, PhD, Mark R Hemmila, MD, FACS, Randall S Burd, MD, PhD, FACS, Melanie Neal, MS, N Clay Mann, PhD, Shahid Shafi, MD, MPH, FACS, David E Clark, MD, MPH, FACS, Sandra Goble, MS, Avery B Nathens, MD, PhD, FACS

Trauma systems and trauma centers have been shown to improve outcomes among seriously injured adults[1-8] and children.[9-13] Previous research also suggests that there is variability in care between trauma centers.[14-17] Differences in patient selection (selection bias), case mix, data quality, geography, and other factors inherent to different injured populations likely contribute in part to this variability. However, variability in the processes and quality of care at different trauma centers can also contribute to outcomes variations among hospitals.

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 and to use this information to improve the quality of trauma care in the United States and Canada.[18] The Trauma Quality Improvement Program expands on a foundation of quality improvement programs already conducted by the ACS, including the NSQIP,[19,20] performance improvement/patient safety, and trauma center verification. 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.[15,18]

The Trauma Quality Improvement Program measures quality through comparative estimates and reporting of mortality, complications, and resource use, after accounting for differences in case mix and important confounders. The Trauma Quality Improvement Program also seeks to understand reasons for variability in trauma care, to learn from high-performing hospitals, and to provide constructive feedback to participating trauma centers that will maximize health outcomes among trauma patients. Although NSQIP has served as an example for TQIP with several similarities, trauma patients and trauma care are distinct from nontrauma surgical patients and their care.[21] These differences require a unique approach to risk-adjusted benchmarking and measurement of quality in trauma. Although previous publications have detailed the inception, vision, and feasibility of TQIP,[15,18] the methodology used for risk adjustment and benchmarking have not been reported.

The objective of this article is to detail the methodology, data processing, data quality, statistical analysis, and analytic rationale for TQIP. A group of experts in trauma research methodology and statistical analysis (the TQIP Analytics Project Team) was tasked with developing the TQIP methodology, which forms the basis for this article. In presenting the methodological framework and statistical rationale behind TQIP, our goal is to provide transparency about the process of risk-adjusted benchmarking of participating trauma centers.

Study design and setting

The Trauma Quality Improvement Program uses a retrospective cohort of trauma patients meeting specific inclusion criteria and cared for in designated and ACS-verified Level I and II hospitals across the United States and Canada. Trauma center participation in TQIP is voluntary, entails the use of existing trauma registry data conforming to specific standards, and requires an annual

Disclosure Information: Nothing to disclose.

Trauma Quality Improvement Program is supported by funding from the American College of Surgeons.

Received June 12, 2012; Revised August 12, 2012; Accepted August 20, 2012.

From the Center for Policy and Research in Emergency Medicine, Department of Emergency Medicine, Oregon Health & Science University, Portland, OR (Newgard), Department of Surgery, University of Nevada, Las Vegas, NV (Fildes), American College of Surgeons, Chicago, IL (Wu, Neal, Goble), Department of Surgery, University of Michigan Health System, Ann Arbor, MI (Hemmila), Center for Clinical and Community Research, Departments of Surgery and Pediatrics, Children's National Medical Center, Washington, DC (Burd), Intermountain Injury Control Research Center, University of Utah, Salt Lake City, UT (Mann), Department of Surgery, University of Texas Southwestern Medical School, Dallas, TX (Shafi), Department of Surgery, Maine Medical Center, Portland, ME (Clark), and Department of Surgery, University of Toronto, Toronto, Ontario, Canada (Nathens).

Correspondence address: Craig D Newgard, MD, MPH, FACEP, Department of Emergency Medicine, Center for Policy and Research in Emergency Medicine, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd, Mail Code CR-114, Portland, OR 97239-3098. email: newgardc@ohsu.edu

ISSN 1072-7515/12/$36.00
http://dx.doi.org/10.1016/j.jamcollsurg.2012.08.017

Abbreviations and Acronyms

ACS = American College of Surgeons
ED = emergency department
IQR = interquartile range
ISS = Injury Severity Score
LOS = length of stay
O/E = observed to expected
TQIP = Trauma Quality Improvement Program

fee to offset the costs of the program. In this publication, we use information from the most recent TQIP database available for analysis (patients admitted in 2010). Currently, there are 143 participating trauma centers (90 Level I hospitals and 53 Level II hospitals), with the number increasing over time. Participating centers represent a variety of regions, hospital types, and geographic locations (Table 1).

Patient population and inclusion criteria

The Trauma Quality Improvement Program uses a broad, heterogeneous group of seriously injured patients, with focused assessment of several distinct subset populations (Table 2). The aggregate TQIP sample includes adults (age 16 years or older) with at least 1 valid trauma ICD-9-CM diagnosis code (800 to 959.9, excluding diagnosis codes for late effects, superficial injuries, and foreign bodies); blunt or penetrating mechanisms of injury; Abbreviated Injury Scale (AIS) score $\geq$3 (Injury Severity Score [ISS] $\geq$9); and non-missing values for emergency department (ED) and hospital discharge dispositions (Table 2). A pre-existing advanced directive to withhold life-sustaining care is an exclusion criterion. Due to variability among hospitals in classifying patients as "dead on arrival," TQIP mortality analyses are performed both including and excluding patients with an ED discharge disposition of "died." The Trauma Quality Improvement Program reports also exclude elderly patients (65 years or older) with an isolated hip fracture[22]; however, these patients are included in elder-specific reports.

Table 1. Hospital Characteristics for Participating Trauma Quality Improvement Program Hospitals (n = 131)

Hospital characteristics	n	%
Trauma Level		
I	85	65
II	46	35
Bed size, n		
<200	6	5
201–400	36	27
401–600	41	31
>600	48	37
Teaching type		
University	65	50
Community teaching	54	41
Community nonteaching	12	9
Hospital type		
For profit	9	7
Nonprofit	122	93
Region		
Northeast	17	13
Midwest	44	34
South	40	31
West	30	23

Based on the number of participating centers with data available at the time of this report.

The Trauma Quality Improvement Program specifies several different cohorts to address different aspects of trauma care. These groups include blunt multisystem injury (AIS $\geq$ 3 in at least 2 body regions); penetrating truncal injury (AIS $\geq$ 3 in the neck, chest or abdomen); shock (systolic blood pressure [SBP] <90 mmHg); isolated traumatic brain injury; and elderly. These cohorts were selected to focus performance and treatment efforts, target distinct types of trauma patients with different needs and management strategies, highlight injury populations with varying representation and experience among centers, and to increase comparability among hospitals. These groups also allow better evaluation of different aspects of multidisciplinary care coordination, timing and strategies of resuscitation, processes of care, expected outcomes, and resource use.

For admissions occurring in 2010, TQIP includes 96,537 trauma patients, 19,586 blunt multisystem injury patients, and 6,440 penetrating injury patients. When assessed on a hospital level, the annual median patient sample size and interquartile range (IQR) are 662 (IQR 409 to 887) total TQIP patients per hospital; 109 (IQR 64 to 194) blunt multisystem patients per hospital; and 35 (IQR 16 to 66) penetrating injury patients per hospital. There is no minimum sample size requirement for a trauma center to participate in TQIP.

Outcomes measures

Primary outcomes include mortality (on arrival, in the ED, and in-hospital), complications and resource use.[23,24] Although in-hospital mortality is influenced by many factors, it is a well-recognized outcome in trauma care, reliably captured in trauma registries and useful for TQIP. For complications, TQIP has focused on addressing potentially preventable events that cause disability, additional resource use, and deviations from the expected clinical course after

Table 2. Inclusion and Exclusion Criteria for Trauma Quality Improvement Program, Plus Criteria for Certain Subset Cohorts

Inclusion criteria	Age 16 y or older At least 1 valid trauma ICD-9 code in the range of 800 to 959.9 (excluding late effects (905−909.9), superficial injuries (910−924.9), and foreign bodies (930−930.9) Primary mechanism of injury classified as either blunt or penetrating: Blunt is defined as an injury where the primary E-code is mapped to the following categories: fall, machinery, motor vehicle traffic, pedestrian, cyclist, and struck by or against Penetrating is defined as an injury where the primary E-code is mapped to the following categories: cut/pierce and firearm Severely injured patients with at least one AIS ≥3 injury: For blunt injuries: at least 1 injury in any of the following AIS body regions: head, face, neck, thorax, abdomen, spine, or upper and lower extremity For penetrating injuries: at least one AIS ≥3 injury in any of the following AIS body regions: neck, thorax, and abdomen Injury severity score ≥9 ED discharge disposition and hospital discharge disposition cannot both be unknown
Exclusion criteria	Comorbidity: pre-existing advanced directive to withhold life-sustaining interventions Isolated hip fractures for patients 65 y or older with an injury with mechanism of fall is defined as any traumatic injury with at least one of the following diagnosis codes: 851810.3 Femur, fracture, intertrochanteric 851812.3 Femur, fracture, neck 851818.3 Femur, fracture, subtrochanteric and all other injuries in AIS body region "external" (ie, bruise, abrasion, or laceration)
Elderly patients without isolated hip fractures	Patients 65 y or older and without isolated hip fractures
Elderly patients with isolated hip fractures	Patients 65 y or older and with isolated hip fractures
Isolated traumatic brain injury patients	Patients met one of the following criteria: AIS severity ≥4 for body region head and no other severe injuries in any other body region or AIS severity ≥3 for body region head and initial Glasgow Coma Scale motor score in ED ≤4 and no other severe injuries in any other body region
Shock patients	Patients with ED SBP of ≤90

AIS, Abbreviated Injury Scale; ED, emergency department; SBP, systolic blood pressure.

injury. Targeted complications include urinary tract infection, deep venous thrombosis, pulmonary embolism, ventilator-associated pneumonia, central line−related bacteremia, renal failure, and surgical site infections. Measures of resource use (eg, length of stay [LOS], duration ICU stay and ventilator days) were selected based on feasibility of data capture, association with quality of care, relation to other TQIP outcomes (eg, complications), responsiveness to evidence-based practice guidelines, and a direct relationship with cost.

Process of care metrics in specific target populations are also used in TQIP to focus on particular quality issues in trauma care. Examples include intracranial pressure monitoring in severe traumatic brain injury; operative timing (eg, time to operative fixation in long-bone fractures); placement and timing of tracheostomy; time to hemorrhage control; and venous thromboembolism prophylaxis. Where possible, TQIP uses process metrics that exist in published guidelines. In cases where outcomes-based evidence is not strong (eg, tracheostomy placement), TQIP reports the comparative practice patterns and timing back to centers without specifying a particular quality target. Even when not supported by evidence of improved outcomes, comparative results allow centers to relate their practices to those in other trauma centers. This comparative assessment can be helpful in providing insight for outlying hospitals and potentially explaining other local findings (eg, complication rates, deviation in LOS).

Data processing and data quality

Compilation of high-quality, reliable data is an integral aspect of TQIP. The program capitalizes on existing trauma registry infrastructure at trauma centers and requires use of the National Trauma Data Standard to assure consistent chart abstraction, data definitions, and information source hierarchy. Due to large variability in inclusion criteria among trauma registries,[25] TQIP uses an injury severity threshold (AIS ≥3) high enough that all registries capture the target population.

There are several data quality assurance mechanisms in TQIP to assure consistent, high-quality data collection.

These mechanisms include training courses for trauma registrars and data abstractors; monthly data quality educational activities (eg, conference calls, quizzes, and webinars); data logic checks; assessment of outlier values; internal validation to verify the appropriateness and completeness of data; and external validation of each hospital's data. External validation includes a site visit to participating hospitals every 3 years to assess processes for case identification, data abstraction, data entry, and data quality (including re-abstraction of a random subset of trauma charts). Site visits also provide an opportunity to learn from individual centers, understand how TQIP reports are used by hospitals, and develop strategies for improving TQIP and the resulting quality improvement efforts.

The Trauma Quality Improvement Program also generates separate data quality reports for each participating hospital to assure reliable, consistent, and accurate data. These data quality reports contain comparative assessments (ie, compared with other participating hospitals) of missing values, case ascertainment, application of inclusion/exclusion criteria, data quality benchmarks, and data fields with considerable variability in quality (eg, comorbid conditions and complications). The data quality reports also evaluate key data fields among prespecified subsets of patients (eg, ISS among patients that die within 1 day of admission; complications among patients with ISS >24 and LOS >1 day) intended to highlight data quality issues.

Variables

Multiple data elements are captured for TQIP and considered in risk-adjustment models. These variables include patient demographics, comorbid conditions, initial ED physiology, ED disposition, transfer status, mechanism of injury, ICD-9-CM diagnosis codes, procedures, AIS scores, derived injury severity measures, LOS, ICU stay, complications, and in-hospital mortality (Table 3).

Injury severity measures are a critical aspect of describing, stratifying, analyzing, and risk-adjusting trauma centers. Hospitals vary in how these measures are captured and calculated. Some hospitals manually abstract hospital records to code AIS values for individual injuries, whereas others use an AIS mapping function that generates AIS values from descriptive injury information. Other hospitals do not directly capture any AIS values. Approximately half of TQIP centers use the AIS-98 format, and the remainder use AIS-05 and older versions of AIS coding (eg, AIS-90). To provide consistent injury severity coding between hospitals, TQIP uses an algorithm to convert AIS-90 codes and the more granular AIS-05 codes to AIS-98 codes. For the few centers that do not provide

Table 3. Variables Considered in Trauma Quality Improvement Program Multivariable Models

Mortality model
Initial GCS motor score in ED
Initial systolic BP in ED
Initial pulse rate in ED
Mechanism of injury
Pedestrian/pedal: motor vehicle-pedal cyclist, motor vehicle-pedestrian, pedal cyclist/other, pedestrian/other
Motor vehicle occupant and other motor vehicle related event
Motorcyclist
Fall
Struck by or against
Firearm
Cut/pierce
Other
Transfer status
Age
Gender
Race and ethnicity
AIS severity by individual body region (except for external)
Individual comorbidities
Heart disease
Cancer
Liver disease
Alcoholism
Smoking
Stroke
Diabetes
Hypertension
Renal disease
Impaired sensorium
Respiratory disease
Functional dependence
Bleeding disorder
Peripheral vascular disease
Steroid use (included if prevalence >2%)
Region (West, South, Midwest, Northeast)
Payment type
Derived variables
Injury Severity Score
ICD9-based Injury Severity Score
SWI (based on ICD9 injury codes)
Maximum AIS by body region
Lowest AIS = lowest AIS score
Serious AIS = maximum AIS ≥3 for specific body regions
Arrest SBP = emergency department SBP ≤40 mmHg
Length of stay model
Same covariates noted above, plus complications
Cardiovascular (cardiac arrest with CPR, myocardial infarction, stroke)
Surgical infections (organ/space surgical site infection, deep surgical site infection, superficial surgical site infection, and wound disruption)
Acute respiratory distress syndrome
Pulmonary embolism
Renal failure
Pneumonia
Sepsis

AIS, Abbreviated Injury Scale; ED, emergency department; SBP, systolic blood pressure; SWI, Single Worst Injury.

AIS codes or that code in a version older than AIS-90, TQIP applies an ICD-9-CM to AIS-98 mapping algorithm to derive AIS codes (ICDmap 90, 1995 update; Windows version, Johns Hopkins University, 1997). Previous studies have validated software for mapping administrative diagnosis codes to anatomic injury scores.[26-28]

Once AIS scores have been generated for all TQIP patients, an ISS is calculated from the AIS values. ICD-9-CM diagnosis codes are also used to generate a separate measure of injury severity, termed the *ICD-9 Injury Severity Score*,[29] which is calculated by first creating survival risk ratios for every ICD-9 injury diagnosis in a reference population. For patients in TQIP, the Single Worst Injury ICD-9 Injury Severity Score is used for modeling. The goal in generating injury severity measures between hospitals is consistency in injury coding and therefore comparability of results between institutions. To improve consistency in injury scoring, TQIP is working with participants to encourage and ultimately require the use of AIS-05 formatting of injury scores, with generation of survival risk ratios for all codes in the AIS-05 lexicon.

Handling missing values

Missing values are frequently present in trauma registries.[30-32] Although the simplest solution for handling missing values is to restrict analyses to patients with observed values (complete case analysis), this approach can introduce bias, reduce sample size, and reduce study power.[30,33-38] Such an approach to handling missing data is a recognized threat to quality improvement efforts and trauma center benchmarking.[31,39] A previous TQIP data quality project compared hospital rankings of observed-to-expected mortality with and without accounting for missing data and found that approximately 20% of hospitals changed their rank (better or worse), depending on how missing values were handled.[40]

To maximize the rigor and validity of TQIP reports, we use multiple imputation[41] to handle missing values. Multiple imputation is an analytic method that uses observed values to generate a range of plausible values for each previously missing data point based on existing correlations and relationships between variables.[33,41] The primary assumption required for multiple imputation is that the mechanism of missingness is either "missing completely at random" (missing values are independent of observed and unobserved covariates) or "missing at random" (missing values are not completely random, but their mechanism can be explained by observed values and is therefore "ignorable").[33,41] The result of multiple imputation is multiple complete datasets with no missing values, each of which is analyzed independently using standard parametric statistical methods. Results for each multiply imputed dataset are combined using standardized rules to appropriately account for the variability within and between datasets and, therefore, the uncertainty inherent in the imputation process.[33,41]

The TQIP imputation models include a group of demographic, clinical, procedural, process, and outcome measures. Variables in the models include age, sex, comorbidities, transfer status, SBP, pulse, Glasgow Coma Scale motor score, Single Worst Injury ICD-9 Injury Severity Score, and maximum AIS severity of each body region. We use PROC MI (SAS v 9.3; SAS Institute) for multiple imputation models, with Markov Chain Monte Carlo methodology to specify multivariable associations between all variables and to generate a posterior probability distribution from which to select the imputed values. Previous research has described the validity of and modeling approach for using multiple imputation to handle missing physiologic data in the National Trauma Data Bank.[32,42] Several additional studies have demonstrated the validity and rigor of multiple imputation for handling missing trauma data (prehospital and in-hospital) under a variety of conditions and across multiple trauma systems.[30,38,43]

Statistical analysis

In-hospital mortality

Providing valid risk-adjusted mortality estimates for trauma center comparison and benchmarking is a central goal of TQIP. Patient populations vary across hospitals by demographics, acuity, mechanism of injury, timing of presentation, and comorbidities. Methods are needed to account for these differences. Multiple approaches to risk-adjusted modeling were considered for TQIP, including logistic regression, hierarchical models, generalized estimating equations, Bayesian analysis, linear regression, and Poisson regression. Although each of these approaches has certain advantages, there were concerns that overly complicated approaches to modeling would reduce the face validity and interpretability of TQIP reports, and potentially create analytic obstacles. For example, hierarchical and generalized estimating equation models might be preferable to account for correlated data (clustering), but fully accounting for such clustering at the hospital level can inadvertently "adjust away" some of the key inter-hospital process and outcomes differences that are the focus of TQIP. Another example is Bayesian analysis, which has advantages in increasing the precision of hospital risk-adjusted estimates, but is not compatible with multiple imputation. After evaluating the merits and limitations of different types of models, we selected multivariable logistic regression for the primary

mortality model based on its face validity (widely recognizable and easily understood), generation of risk-adjusted estimates that preserve center-level differences as quality targets, compatibility with multiple imputation and equivalent performance with less complexity for risk-adjusting trauma care.

Several multivariable logistic regression models were developed and tested before deciding on the final TQIP risk-adjusted mortality model. The initial model was built using all potential covariates and confounders (Table 4). Continuous variables (eg, SBP and pulse) were entered into the model separately as either continuous or categorical (eg, SBP <90 mmHg) terms to allow for the best model discrimination and fit and to keep the models parsimonious. We also considered covariates for geographic region, availability of discharge options (eg, rehabilitation facilities, skilled nursing facilities) and insurance status. Table 4 shows a comparison of multivariable risk-adjusted mortality models considered for TQIP using a sample of 18,444 patients. The final TQIP risk-adjustment mortality model includes 18 variables and was selected based on goodness-of-fit, model discrimination, predicted vs observed mortality across deciles of risk (Fig. 1), inclusion of important confounders, and iterative model building techniques. The goal was to create the most parsimonious model that maintained stable covariate point estimates (direction and magnitude of effect), statistical associations, and model performance (described later). Because each cohort within TQIP has unique features and different associations with predictor variables, the models differ slightly among the cohorts. However, the modeling strategy, variables considered in the models, and assessment of model performance are the same across all cohorts.

Table 4. Comparison of Multivariable Logistic Regression Models for Risk-Adjusted Mortality in Trauma Quality Improvement Program (n = 18,444)

	Model 1	Model 2	Model 3 (final)
Model type	Logistic	Logistic	Logistic
No. of variables	31	17	18
Model performance			
Hosmer-Lemeshow goodness-of-fit p value*	0.04	0.08	0.09
c-statistic†	0.90	0.90	0.90
AIC‡	7,945.2	7,940.9	7,935.6

Model 1 included age, sex, injury mechanism, transfer status, race, systolic blood pressure (SBP), arrest SBP, Glasgow Coma Scale (GCS) motor score, pulse, Single Worst Injury (SWI), head Abbreviated Injury Scale (AIS), neck AIS, chest AIS, abdominal AIS, spine AIS, lower AIS, heart disease, cancer, liver disease, alcohol, smoker, stroke, diabetes, hypertension, dialysis, impaired sensorium, obesity, respiratory disease, functional dependence, bleeding disorder, and peripheral vascular disease. Model 2 included age, injury mechanism, transfer status, SBP, GCS motor score, pulse, SWI, head AIS, lower AIS, heart disease, cancer, liver disease, hypertension, dialysis, impaired sensorium, functional dependence, bleeding disorder, and peripheral vascular disease. Model 3 included Model 2 + arrest SBP.

*For the Hosmer-Lemeshow goodness of fit statistic, a p value >0.05 indicates better model 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.

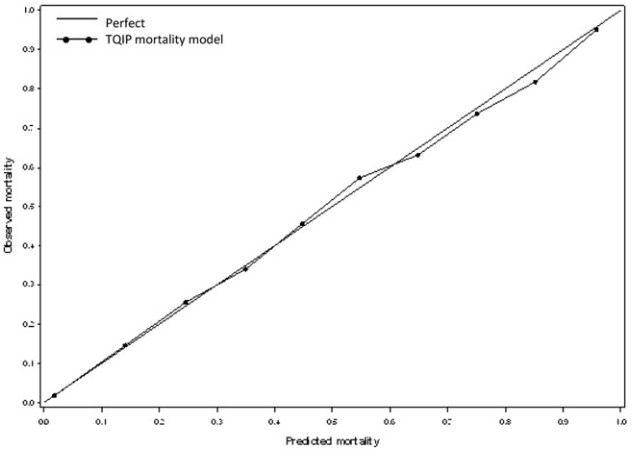

Figure 1. Assessment of Trauma Quality Improvement Program model performance (n = 85,569).

Model performance

Model performance was an integral factor in selecting the final models for TQIP. Performance metrics included the Hosmer-Lemeshow goodness-of-fit statistic (calibration), calibration curves, c-statistic (discrimination), and Akaike information criterion value (to compare model fit and composition across multiple models). After inclusion of relevant covariates, a stepwise variable selection process was used to generate the most parsimonious model, evaluate changes in confidence intervals, and examine shifts in the magnitude and direction of point estimates. Risk factors were "forced" into the risk-adjustment model if they were considered important confounders for clinical outcomes. Because no one metric can provide comprehensive information on model performance, all metrics were examined throughout the process. As the Hosmer-Lemeshow goodness-of-fit statistic can be overly sensitive in identifying poorly fit models with large sample sizes, we supplemented this statistic with calibration curves to visually represent predicted vs observed events from the regression model (Fig. 1). The c-statistic, a common tool used to evaluate a model's ability to discriminate between patients with and without outcomes events (eg, deaths),[44] was typically ≥85% for TQIP models. Akaike information criterion values were used to compare the fit and composition of different TQIP models.

Observed-to-expected ratios and the W-statistic

The W-statistic has been used to compare observed to expected deaths among injured patients. This practice began with the Major Trauma Outcomes Study, in which the Trauma and Injury Severity Score (TRISS) was used to adjust for differences in case mix across centers.[45] The Trauma and Injury Severity Score provided probabilities of survival at the patient level using relatively few variables for risk adjustment. The W-statistic was then calculated using the observed minus expected outcomes differences to estimate the number of excess deaths (or survivors) per 100 patients. Unfortunately, the regression coefficients for TRISS were derived in the 1980s rather than using outcomes associated with more contemporary trauma care, and TRISS methodology has been questioned for trauma center benchmarking.[46]

Although the W-statistic can be estimated using TQIP risk-adjustment models, we have elected to use the observed-to-expected (O/E) ratio. The O/E ratio has its foundations in the more generic standardized mortality ratio used in other areas of medicine. Additionally, stakeholders from a wide variety of backgrounds have become familiar with the O/E ratio through other quality improvement programs, such as NSQIP. The O/E ratio is calculated using the models described here to estimate an expected outcomes "risk" (eg, probability of mortality between 0 and 1) for each patient included in the sample, adjusted for all covariates in the model. The expected number of outcomes at a given trauma center is then generated by summing all probability values for patients treated at the hospital to provide a risk-adjusted estimate for the number of expected outcomes events. The number of actual (observed) outcomes is divided by the number of expected outcome to yield the O/E ratio for each participating hospital by year. Because the precision of these estimates varies by sample size, patient characteristics, model performance and missing data, it is important to also provide estimates of variance (confidence) for the O/E ratios. For TQIP, 95% CI are calculated for the O/E estimates using a Bernoulli distribution approximation method for observed outcomes. A 90% CI is used in certain analyses for cohorts with small sample sizes (eg, penetrating injury cohort).

Length of stay

Similar models were developed to produce risk-adjusted estimates for LOS as a measure of resource use. Risk factors similar to those described for the mortality model were considered in the LOS model (Table 3). Because LOS can be affected by in-hospital mortality rates (eg, a hospital with high in-hospital mortality may appear to have low LOS), we opted to restrict LOS models to survivors for simplicity and clarity. Patients dying during their stay represent a competitive risk, whereby death "competes" with LOS by shortening the duration of hospital stay. Also, LOS is a continuous (or count) variable with a non-normal, right-skewed distribution, requiring consideration of a variety of model types, such as linear regression (after normalization of LOS), Poisson, negative binomial, and gamma distribution models. Due to overdispersion of the data, TQIP currently uses a zero-truncated negative binomial model to predict LOS based on the same predictor variables used in the mortality model. The LOS model also includes the presence of complications (eg, cardiovascular, surgical infections, acute respiratory distress syndrome, pulmonary embolism, renal failure, pneumonia, and sepsis), payer type and region (a surrogate for rehabilitation and skilled nursing home availability). The proportion of patients having "excess" LOS is calculated for each center, defined as the proportion of patients with observed LOS 25% greater than predicted or observed LOS in the $\geq$95% ile. A 95% CI for the proportion of excess LOS for each trauma center is calculated using Clopper-Pearson methodology[47] to compare LOS performance across TQIP hospitals. The LOS models used for TQIP differ by patient cohort, as fitting the distribution for LOS changes substantially for different injured populations.

Generating accurate, risk-adjusted estimates for LOS in an easy-to-understand format remains both a goal and a challenge in TQIP. There are many factors that contribute to variation in LOS, which leads to difficulty in accurately modeling this term. The Trauma Quality Improvement Program continues to explore different models, modeling strategies, and additional covariates to improve the LOS model predictability and fit, as well as the value of TQIP reports.

Other outcomes models

Models are being developed for complications and additional measures of resources use, as well as for the different subset cohorts. For these models, available sample size and number of outcomes at each hospital are major factors in determining model specifications, stability of the estimates, and statistical feasibility. As TQIP is a work in progress, additional models and key subset cohorts will continue to be refined based on feedback from participating hospitals and the need to address specific clinical questions to maximize the quality of trauma care.

Presentation of results

One of the key goals in TQIP is to provide readily interpretable and informative comparisons of trauma center performance. Initial TQIP reports used rank plots (also

termed *caterpillar* plots),[15] which provide a familiar and somewhat traditional method for comparative "ranking" of hospitals (Fig. 2A). However, these plots have important limitations, including a rank-order list that is not necessarily meaningful (eg, in Fig. 2A, sites "I" through "R" could be ranked 9th through 18th, but without statistical difference in rank); the resulting potential for misinterpretation; lack of hospital sample size information (inability to compare outcomes among similar-volume hospitals); and difficulty in differentiating hospitals that are close to outlier status. For these reasons, TQIP also uses funnel plots (Fig. 2B). Funnel plots allow direct assessment of trauma center volume, improved visual assessment of outlier hospitals (high and low), elimination of nonmeaningful hospital rankings, and easier identification of hospitals close to outlier status (eg, early recognition of quality issues that can prompt behavior change, even if not yet statistically significant).[48,49]

There are differences when comparing the 2 strategies for visual presentation of TQIP data, reflecting different

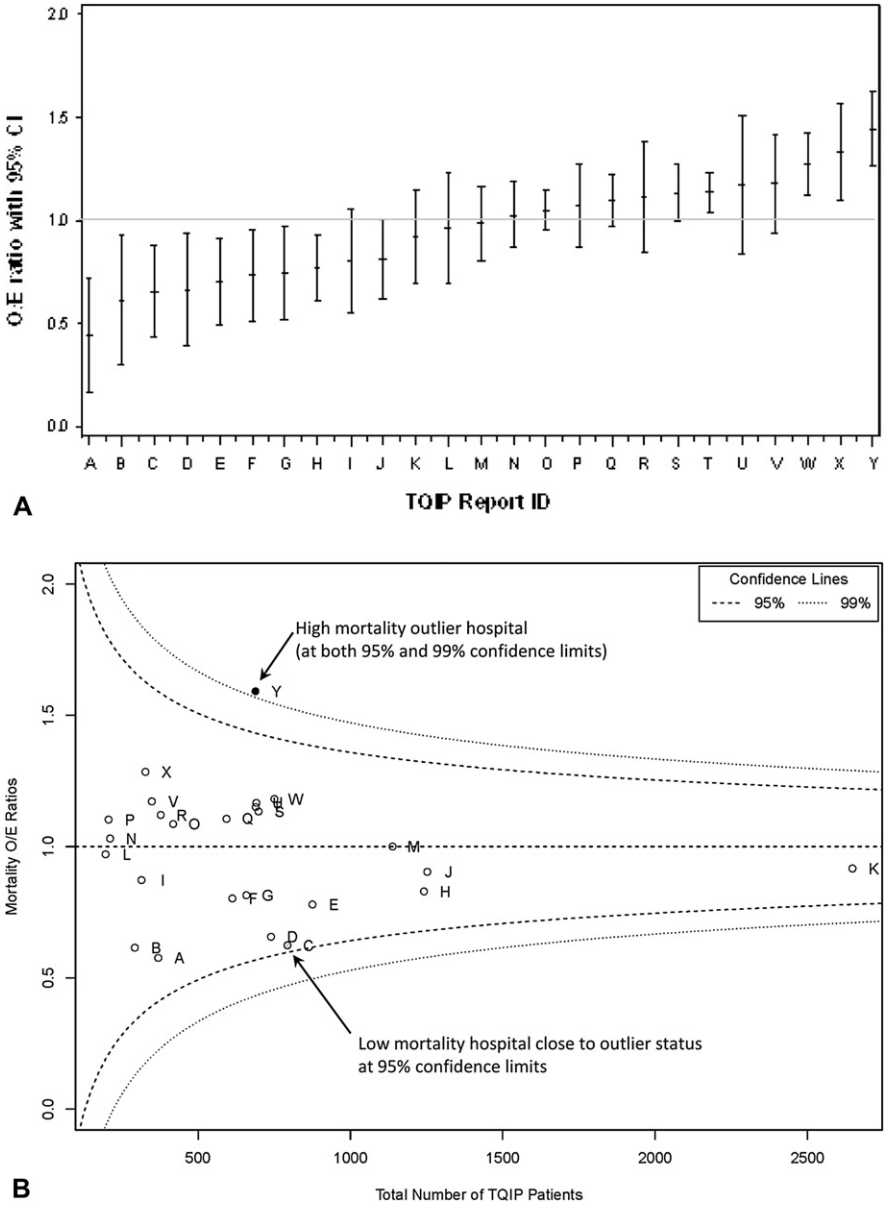

Figure 2. Visual depiction of risk-adjusted observed-to-expected mortality ratios for 25 randomly selected trauma centers (from a sample of 85,569 patients treated in 113 trauma centers). (A) Rank (caterpillar) plot. (B) Funnel plot.

statistical techniques for assessing outlier hospitals and handling variance. The rank plot (Fig. 2A) identifies 12 outlier hospitals at the 95% confidence level, and the funnel plot (Fig. 2B) identifies one outlier and one site close to outlier status. In the funnel plot, hospital Y has a statistically higher than expected adjusted mortality rate (at 95% and 99% confidence levels). Other high outliers on the rank plot are comparable with the other hospitals in the funnel plot. Conversely, hospital C is close to the 95% CI line on the funnel plot (statistically close to low mortality outlier status), although other low outliers from the rank plot are within the range of error (variance) in the funnel plot. These statistical methods are being refined to produce consistent results that balance the sensitivity for identifying outliers with an appropriate level of confidence. The Trauma Quality Improvement Program will continue to use both reporting formats (rank plots and funnel plots), as participants have found value in both types of figures for understanding their data.

Limitations

As with any data-driven quality improvement effort, the value of TQIP is highly dependent on using high-quality data and analytically sound methodology for generating risk-adjusted estimates and benchmarking. The data quality program and statistical methodologies were selected to meet such specifications, although there are limitations to consider. First, complete case ascertainment is critically important to ensure that all eligible patients meeting the TQIP inclusion criteria at a given trauma center are included in the TQIP estimates and to minimize selection bias. Complete ascertainment requires regular review of ED and trauma patient logs for eligible patients and confirming the completeness of relevant trauma registry data fields for inclusion criteria (eg, AIS scores). Although capturing all eligible patients might seem straightforward, some research suggests that key data fields can be systematically biased or missing in certain groups of high-risk trauma patients (eg, inappropriately low ISS values in patients with early death, where diagnostic testing or autopsy were not completed).[50] Next, the quality of data from participating centers is crucial. Without high-quality primary data collection at the hospital-level, subsequent results can be biased. The Trauma Quality Improvement Program has implemented a comprehensive data quality program to maximize the likelihood of receiving high-quality data, including data standards, quality assurance mechanisms, and external data validation.

The use of multiple imputation requires that the mechanism of missingness is ignorable.[33,41] Based on previous studies evaluating the validity of multiple imputation for trauma registry data, the National Trauma Data Bank, and other forms of trauma data,[30,32,38,42,43] we believe this assumption holds for TQIP. TQIP models are based on available data fields to account for confounding and differences in case mix, however, there is the potential that unmeasured confounding (bias) could alter results. In addition, the timeliness of TQIP data and feedback reports has a lag time directly related to the receipt of trauma registry data from sites.

Finally, the precision of TQIP estimates is highly dependent on sample size, which can be low for certain hospitals when evaluating highly selected populations (eg, penetrating injury). This lack of precision might suggest that a given hospital is providing care consistent with that of other trauma centers, although estimates with greater precision could suggest otherwise. We have tried to develop statistical models that maximize statistical efficiency and precision, although sample size will remain a limitation in identifying certain outlier hospitals. The ability to "drill down" in TQIP data to address specific quality improvement questions might therefore be limited by center-based sample sizes.

Potential applications and impact

There are several applications of TQIP reports and potential for impact. Although use of the TQIP reports will likely differ between hospitals, this comparative feedback should serve as a constructive tool to drive trauma quality improvement efforts at the hospital level. The TQIP results offer an opportunity to compare trauma center outcomes and processes in a way that identifies aspects of trauma center care that are working well vs those that need attention. The TQIP reports have been likened to "a warning light on the dashboard," requiring individual sites to "look under the hood" at their own institutions to evaluate the nature of the problem and potential solutions. Ideally, TQIP feedback reports will reduce variability in trauma care, complications, unnecessarily long hospital stays, and improve survival. However, realizing these goals will take time and effort.

Early feedback from participating hospitals suggests that TQIP reports have already been used in a variety of ways. Such applications have included strengthening process improvement efforts, improving the quality of registry data, increasing staffing levels, and targeting specific types of trauma care (eg, elder care) where comparison with peer centers suggested the need for improvement. In some states, the cost of TQIP has been borne to a large extent by a principal insurance payer to improve the quality of hospital care. Although verification and state designation processes examine structure and certain aspects of quality, these efforts do not compare hospital-specific, risk-adjusted trauma outcomes

between centers. The Trauma Quality Improvement Program provides a direct feedback loop to trauma centers about national risk-adjusted performance and the opportunity to learn from high-performing centers by sharing best practices. The Trauma Quality Improvement Program is not intended to be a punitive program, but rather one that allows focused self-assessment at the hospital level and data-driven decisions toward providing higher-quality trauma care.

CONCLUSIONS

The Trauma Quality Improvement Program is a national trauma quality improvement effort based on rigorous analytic methodology for risk-adjusted benchmarking of trauma centers. Case ascertainment, data quality, handling of missing values, statistical modeling, model performance, and presentation/feedback of TQIP results to participating trauma centers is intended to maximize the usability and impact of TQIP findings. The Trauma Quality Improvement Program offers the potential for significant impact in reducing variability and improving the quality of care among US and Canadian trauma centers.

Author Contributions

Study conception and design: Newgard, Fildes, Wu, Hemmila, Burd, Neal, Mann, Shafi, Clark, Goble, Nathens

Acquisition of data: Wu, Neal, Goble, Nathens

Analysis and interpretation of data: Newgard, Wu, Hemmila, Burd, Neal, Mann, Shafi, Clark, Goble, Nathens

Drafting of manuscript: Newgard

Critical revision: Newgard, Fildes, Wu, Hemmila, Burd, Neal, Mann, Shafi, Clark, Goble, Nathens

REFERENCES

1. MacKenzie EJ, Rivara FP, Jurkovich GJ, et al. A national evaluation of the effect of trauma-center care on mortality. N Engl J Med 2006;354:366–378.
2. Mullins RJ, Veum-Stone J, Helfand M, et al. Outcome of hospitalized injured patients after institution of a trauma system in an urban area. JAMA 1994;271:1919–1924.
3. Sampalis JS, Denis R, Lavoie A, et al. Trauma care regionalization: a process-outcome evaluation. J Trauma 1999;46: 565–581.
4. Mullins RJ, Veum-Stone J, Hedges JR, et al. Influence of a statewide trauma system on location of hospitalization and outcome of injured patients. J Trauma 1996;40: 536–545.
5. Pracht EE, Tepas JJ, Celso BG, et al. Survival advantage associated with treatment of injury at designated trauma centers. Med Care Res Rev 2007;64:83–97.
6. Nathens AB, Jurkovich GJ, Rivara FP, Maier RV. Effectiveness of state trauma systems in reducing injury-related mortality: a national evaluation. J Trauma 2000;48:25–30.
7. Mullins RJ, Mann NC. Population-based research assessing the effectiveness of trauma systems. J Trauma 1999; 47[Suppl]:S59–S66.
8. Jurkovich GJ, Mock C. Systematic review of trauma system effectiveness based on registry comparisons. J Trauma 1999; 47[Suppl]:S46–S55.
9. Cooper A, Barlow B, DiScala C, et al. Efficacy of pediatric trauma care: results of a population-based study. J Pediatr Surg 1993;28:299–303.
10. Hulka F, Mullins RJ, Mann NC, et al. Influence of a statewide trauma system on pediatric hospitalization and outcome. J Trauma 1997;42:514–519.
11. Johnson DL, Krishnamurthy S. Send severely head-injured children to a pediatric trauma center. Pediatr Neurosurg 1996;25:309–314.
12. Hall JR, Reyes HM, Meller JL, et al. The outcome for children with blunt trauma is best at a pediatric trauma center. J Pediatr Surg 1996;31:72–77.
13. Pracht EE, Tepas JJ, Langland-Orban B, et al. Do pediatric patients with trauma in Florida have reduced mortality rates when treated in designated trauma centers? J Pediatr Surg 2008;43:212–221.
14. Cudnik MT, Sayre MR, Hiestand B, Steinberg SM. Are all trauma centers created equally? A statewide analysis. Acad Emerg Med 2010;17:701–708.
15. Hemmila MR, Nathens AB, Shafi S, et al. The Trauma Quality Improvement Program: pilot study and initial demonstration of feasibility. J Trauma 2010;68:253–262.
16. Trooskin SZ, Copes WS, Bain LW, et al. Variability in trauma center outcomes for patients with moderate intracranial injury. J Trauma 2004;57:998–1005.
17. Shafi S, Nathens AB, Parks J, et al. Trauma quality improvement using risk-adjusted outcomes. J Trauma 2008;64:599–604.
18. Shafi S, Nathens AB, Cryer HG, et al. The Trauma Quality Improvement Program of the American College of Surgeons Committee on Trauma. J Am Coll Surg 2009;209:521–530.
19. Birkmeyer JD, Shahian DM, Dimick JB, et al. Blueprint for a new American College of Surgeons: National Surgical Quality Improvement Program. J Am Coll Surg 2008;207: 777–782.
20. Fink AS, Campbell DA Jr, Mentzer RM Jr, et al. The National Surgical Quality Improvement Program in non-Veterans Administration hospitals: initial demonstration of feasibility. Ann Surg 2002;236:344–353.
21. Hemmila MR, Jakubus JL, Wahl WL, et al. Detecting the blind spot: complications in the trauma registry and trauma quality improvement. Surgery 2007;142:439–448.
22. Gomez D, Haas B, Hemmila M, et al. Hips can lie: impact of excluding isolated hip fractures on external benchmarking of trauma center performance. J Trauma 2010;69:1037–1041.
23. Ingraham AM, Xiong W, Hemmila MR, et al. The attributable mortality and length of stay of trauma-related complications: a matched cohort study. Ann Surg 2010;252:358–362.
24. Shafi S, Barnes S, Nicewander D, et al. Health care reform at trauma centers—mortality, complications and length of stay. J Trauma 2010;69:1367–1371.
25. Mann NC, Guice K, Cassidy L, et al. Are statewide trauma registries comparable? Reaching for a national trauma dataset. Acad Emerg Med 2006;13:946–953.
26. Clark DE, Osler TM, Hahn DR. ICDPIC: Stata Module to Provide Methods for Translating International Classification of Diseases (Ninth Revision) Diagnosis Codes into Standard

Injury Categories and/or Scores. Boston: Boston College, Department of Economics; 2009.
27. MacKenzie EJ, Steinwachs DM, Shankar BS, Turney SZ. An ICD-9CM to AIS conversion table: development and application. Proc AAAM 1986;30:135−151.
28. MacKenzie EJ, Steinwachs DM, Shankar B. Classifying trauma severity based on hospital discharge diagnoses. Validation of an ICD-9CM to AIS-85 conversion table. Med Care 1989;27:412−422.
29. Osler T, Rutledge R, Deis J, Bedrick E. ICISS: an International Classification of Disease-9 based Injury Severity Score. J Trauma 1996;41:380−386; discussion 386−388.
30. Newgard CD. The validity of using multiple imputation for missing prehospital data in a state trauma registry. Acad Emerg Med 2006;13:314−324.
31. O'Reilly G, Jolley D, Cameron P, Gabbe B. Missing in action: a case study of the application of methods for dealing with missing data to trauma system benchmarking. Acad Emerg Med 2010;17:1122−1129.
32. Moore L, Hanley JA, Turgeon AF, et al. A multiple imputation model for imputing missing physiologic data in the national trauma data bank. J Am Coll Surg 2009;209:572−579.
33. Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. New York, John Wiley & Sons, Inc.; 2002.
34. Van Der Heijden GJMG, Donders ART, Stijnen T, Moons KGM. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006;59:1102−1109.
35. Crawford SL, Tennstedt SL, McKinlay JB. A comparison of analytic methods for non-random missingness of outcome data. J Clin Epidemiol 1995;48:209−219.
36. Joseph L, Belisle P, Tamim H, Sampalis JS. Selection bias found in interpreting analyses with missing data for the prehospital index for trauma. J Clin Epidemiol 2004;57:147−153.
37. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142:1255−1264.
38. Newgard CD, Haukoos J. Missing data in clinical research−part 2: multiple imputation. Acad Emerg Med 2007;14:669−678.
39. Newgard CD, Haukoos JS. Measuring quality with missing data: the invisible threat to national quality initiatives. Acad Emerg Med 2010;17:1130−1133.
40. Glance LG, Osler TM, Mukamel DB, et al. Impact of statistical approaches for handling missing data on trauma center quality. Ann Surg 2009;249:143−148.
41. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York, John Wiley & Sons, Inc.; 1987.
42. Moore L, Hanley JA, Lavoie A, Turgeon A. Evaluating the validity of multiple imputation for missing physiologic data in the national trauma data bank. J Emerg Trauma Shock 2009;2:73−79.
43. Newgard CD, Malveau S, Staudenmayer K, et al. Evaluating the use of existing data sources, probabilistic linkage and multiple imputation to build population-based injury databases across phases of trauma care. Acad Emerg Med 2012; 19:469−480.
44. Merkow RP, Hall BL, Cohen ME, et al. Relevance of the C-statistic when evaluating risk-adjustment models in surgery. J Am Coll Surg 2012;214:822−830.
45. Champion HR, Copes WS, Sacco WJ, et al. The Major Trauma Outcome Study: establishing national norms for trauma care. J Trauma 1990;30:1356−1365.
46. Demetriades D, Chan L, Velmanos GV, et al. TRISS methodology: an inappropriate tool for comparing outcomes between trauma centers. J Am Coll Surg 2001;193:250−254.
47. Clopper C, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26:404−413.
48. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Stat Med 2005;24:1185−1202.
49. Spiegelhalter DJ. Handling over-dispersion of performance indicators. Qual Saf Health Care 2005;14:347−351.
50. Newgard CD, Hedges JR, Diggs B, Mullins RJ. Establishing the need for trauma center care: anatomic injury or resource use? Prehosp Emerg Care 2008;12:451−458.