Pull back the curtain: External data validation is an essential element of quality improvement benchmark reporting

Jill Lynn Jakubus, PA-C, MHSA, MS, Shauna L. Di Pasquo, RN, BSN, Judy N. Mikhail, RN, PhD, Anne H. Cain-Nielsen, MS, Peter C. Jenkins, MD, MSc, *and* **Mark R. Hemmila, MD***, Ann Arbor, Michigan*

BACKGROUND: Accurate and reliable data are pivotal to credible risk-adjusted modeling and hospital benchmarking. Evidence assessing the reliability and accuracy of data elements considered as variables in risk-adjustment modeling and measurement of outcomes is lacking. This deficiency holds the potential to compromise benchmarking integrity. We detail the findings of a longitudinal program to evaluate the impact of external data validation on data validity and reliability for variables utilized in benchmarking of trauma centers.

METHODS: A collaborative quality initiative-based study was conducted of 29 trauma centers from March 2010 through December 2018. Case selection criteria were applied to identify high-yield cases that were likely to challenge data abstractors. There were 127,238 total variables validated (i.e., reabstracted, compared, and reported to trauma centers). Study endpoints included data accuracy (agreement between registry data and contemporaneous documentation) and reliability (consistency of accuracy within and between hospitals). Data accuracy was assessed by mean error rate and type (under capture, inaccurate capture, or over capture). Cohen's kappa estimates were calculated to evaluate reliability.

RESULTS: There were 185,120 patients that met the collaborative inclusion criteria. There were 1,243 submissions reabstracted. The initial validation visit demonstrated the highest mean error rate at 6.2% ± 4.7%, and subsequent validation visits demonstrated a statistically significant decrease in error rate compared with the first visit ($p < 0.05$). The mean hospital error rate within the collaborative steadily improved over time (2010, 8.0%; 2018, 3.2%) compared with the first year ($p < 0.05$). Reliability of substantial or higher (kappa ≥0.61) was demonstrated in 90% of the 20 comorbid conditions considered in the benchmark risk-adjustment modeling, 39% of these variables exhibited a statistically significant ($p < 0.05$) interval decrease in error rate from the initial visit.

CONCLUSION: Implementation of an external data validation program is correlated with increased data accuracy and reliability. Improved data reliability both within and between trauma centers improved risk-adjustment model validity and quality improvement program feedback. (*J Trauma Acute Care Surg.* 2020;89: 199–207.

KEY WORDS: Trauma outcomes; quality improvement; data validation; interrater reliability; trauma registry.

Credible benchmarking of clinical outcomes is reliant upon two essential items: quality data collection and effective risk-adjustment models. Quality data collection, resulting in believable data, is analogous to the construction of a solid foundation and the use of high-grade materials when building a house. Likewise, reliable risk-adjustment modeling represents skilled craftsmanship and is measured by model calibration and discrimination. A considerable amount of literature has addressed the distinctions of different risk-adjustment modeling strategies for national benchmarking of trauma outcomes in the peer review literature.[1–5] Regarding data validity, the American College of Surgeons Trauma Quality Improvement Program (ACS TQIP) has described a vigorous program of training courses, monthly data quality educational activities, data logic checks, data quality reports, and external site visits to validate data at participating hospitals every 3 years.[6,7] Despite these substantial investments in data quality, no critical study of these validation processes or results have been published.

The verification review committee of the ACS Committee on Trauma (ACS-COT) has required risk-adjusted benchmarking as a component of the trauma center verification process since 2016.[8] Utilization of accurate and reliable data is fundamental to creating believable risk-adjusted hospital benchmarking reports. The verification review committee recognizes that the information provided by a trauma registry is only as valid as the data entered and considers monitoring of data validity to be essential. Trauma registry software vendors provide edit checks to detect and flag possible errors in data entry or coding. However, individual trauma centers are primarily tasked with ensuring that the data entered into the trauma registry are accurate and are responsible for conducting their own interrater reliability audits. Still, there remains a scarcity of information in the literature on conducting these audits to guide trauma centers.

Within the Michigan Trauma Quality Improvement Program (MTQIP), we sought to examine the long-term accuracy and reliability of variables considered in the risk-adjusted benchmarking of trauma centers via conduct of an external data validation program. We also investigated the error rates of the

Submitted: August 23, 2019, Revised: October 22, 2019, Accepted: December 1, 2019, Published online: January 7, 2020.

From the Department of Surgery (J.L.J., J.N.M., A.H.C.-N., M.R.H.), University of Michigan, Ann Arbor; Department of Surgery (S.L.D.P.), Beaumont Hospital-Farmington Hills, Farmington Hills, Michigan; and Department of Surgery (P.C.J.), Indiana University, Indianapolis, Indiana.

This study is being presented at the 33rd EAST Annual Scientific Assembly Meeting, January 14–18, 2020, in Orlando, Florida.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text, and links to the digital files are provided in the HTML text of this article on the journal's Web site (www.jtrauma.com).

Address for reprints: Jill L. Jakubus, PA-C, MHSA, MS, University of Michigan NCRC, North Campus Research Complex, Rm 142E, Building 16, 2800 Plymouth Rd, Ann Arbor, MI 48109-2800; email: jjakubus@med.umich.edu.

DOI: 10.1097/TA.0000000000002579

outcomes being measured. In this context, data validation is the process of assessing the quality of already abstracted and submitted trauma registry data by comparison to independently reabstracted data from the electronic medical record (EMR) using standardized data definitions as a reference. We used a data validation and interrater reliability audit approach developed by the ACS National Surgical Quality Improvement Program (NSQIP) and adapted it to the trauma setting.[9–11] In this study, we aim to characterize the influence of this standardized data validation approach on error rate, error by type, and reliability over time within MTQIP.

METHODS

We performed a retrospective cohort study of prospectively collected data. The exposure variable was hospital participation in the MTQIP data validation program, quantified by the number of site visits and years participating in the program. Study endpoints included data accuracy (agreement between registry data and contemporaneous documentation) and reliability (consistency of accuracy within and between hospitals).

Data Sources

A detailed description of the MTQIP structure and components is published.[11] In brief, MTQIP is a collaborative of ACS-COT verified Level I and II trauma centers. The Michigan Trauma Quality Improvement Program began in 2008 as a pilot project with six trauma centers in Michigan. In 2011, MTQIP was accepted as a formal collaborative quality initiative (CQI) sponsored by Blue Cross Blue Shield of Michigan. The program has expanded progressively and currently includes a total of 35 trauma centers. The Michigan Trauma Quality Improvement Program receives data directly from the trauma registry at participant trauma centers via secure file transmission portals at 2-month intervals. The MTQIP members participate in required yearly data validation visits to measure data quality and identify opportunities for system improvement. The total administrative cost of the CQI program, borne by Blue Cross Blue Shield of Michigan, is US $26,000 per center. The cost attributed to data validation is US $2,000 per center.

Data Collection

The Michigan Trauma Quality Improvement Program publishes a data dictionary annually on the collaborative's website (www.mtqip.org). The document consists of National Trauma Data Standard variables as well as MTQIP-specific variables intended to measure processes of care and outcomes. New dictionary iterations were initially accompanied by a webinar reviewing all changes and highlighting potential pitfalls. In 2015, the review session was replaced by a yearly video review available on the MTQIP website. A document is also provided that compares the previous with the upcoming year's data definitions head-to-head and highlighting changes. The data dictionary's change log tracks all historical changes over time. An online orientation video functions as a standardized training instrument for newly enrolled MTQIP participants and new trauma program staff. Data abstraction staff from member trauma centers also attend an annual meeting focused on data quality. Deidentified data validation results are shared, and abstraction practices are discussed in this forum.

Participants are held accountable for the definitions in the MTQIP data dictionary. Data definitions identified as being ambiguous by abstractors are reviewed and clarified in the data dictionary to enhance collaborative transparency and data consistency. Blue font denotes these clarifications in the dictionary. The MTQIP data dictionary provides the validation ranges of data acceptability under each definition (e.g., emergency department [ED] discharge time validation range, ±1 hour).

Data Feedback

The Michigan Trauma Quality Improvement Program uses several feedback mechanisms to highlight potential data quality issues for participants. Registry vendors create edit checks within the registry software to provide data quality feedback before record closure and submission. Upon submission, MTQIP analysts check data fidelity for missing values, inconsistencies, and errors. Participants are notified when issues are encountered and offered an opportunity for correction and resubmission. The MTQIP online data analytics tool includes feedback on the presence of missing values to ensure that the presented outcome does not mislead users. Users can then drill down to create a list of patients with missing or negative date/time values to remedy.

Cohort

Participants submit all trauma registry cases to MTQIP. The Michigan Trauma Quality Improvement Program applies the National Trauma Data Standard trauma registry inclusion criteria to ensure consistent cohort formation. The following MTQIP inclusion criteria are then applied:

- Age ≥ 16 years
- Calculated Injury Severity Score (ISS) ≥ 5
- Primary method of injury classified as either blunt or penetrating:
  - Blunt is defined as an injury where the primary E-code is mapped to the categories of fall, machinery, motor vehicle traffic, pedestrian, cyclist, and struck by against
  - Penetrating is defined as an injury where the primary E-code is mapped to the categories of cut/pierce and firearm
- ED and hospital discharge disposition are not missing
- Length of stay is ≥1 day for patients discharged alive

For data validation, MTQIP applies the case selection criteria provided in Table 1 for all cases submitted with either the disposition of death or cases who were admitted to the trauma, orthopedic surgery, or neurosurgery services. Patients who arrived with no signs of life were excluded from case selection.[12] These criteria were developed, based upon expert opinion, to select for high yield cases that were likely to challenge data abstractors. The criteria logic selects for cases where the probability of an error may be increased based upon a known incidence. The method of pure random selection from all available cases was abandoned because it lacked sufficient sensitivity and led to the predominant selection of low yield cases that were unlikely to have data abstraction issues (e.g., low complexity, short-stay patient).

A one-year interval is queried using the most recently submitted and sanitized data. The selected cases from the 11 criteria are merged. If greater than 10 cases exist for a given criterion, then 10 cases are selected at random before merging. The merged list is then randomized and used to generate a list of

TABLE 1. Case Selection Criteria

Criteria	Criteria Description
1	ISS <16 and mortality
2	ISS >24 and no complications and hospital days >1
3	Length of stay >14 days and no complication or mortality
4	Age > 64 and no comorbidities
5	Mechanical ventilator days >7 and no pneumonia
6	Motor Glasgow Coma Score = 1 and no complications and hospital days >1
7	Hematocrit <22.0 and no transfusion of packed red blood cells within 4 hours or IV fluid administration captured
8	ISS >24 and no complications and intensive care unit days >7
9	ISS >9 and no injury in the abbreviated injury scale head and no venous thromboembolism prophylaxis and length of stay >2 days
10	ED blood pressure <90 mm Hg and lowest systolic blood pressure <90 mm Hg and transfusion of packed blood cells within 4 h = 0
11	Antibiotic days >6 and no complications

IV, intravenous; mm Hg, millimeters of mercury.

25 cases for potential data validation. From this list, 10 randomly selected cases are highlighted and submitted to the center, so these records can be made available for data validation.

The selection of data elements on which to perform data validation has evolved. Variables that were not being used in the risk-adjustment modeling or have demonstrated consistent capture based on error rates were removed from the validation process over time. New variables have been included in the validation process as data elements are added and used to conduct expanded quality improvement efforts targeted by MTQIP.

Rater Selection

Each case is reviewed by the MTQIP program manager and registrar validator. The program manager completed the ACS NSQIP Surgical Clinical Reviewer Training, Association for the Advancement of Automotive Medicine (AAAM) Abbreviated Injury Scale (AIS) training, emergency medical technician training, and physician assistant school. She is employed as a physician assistant in acute care surgery at an ACS-COT verified Level I trauma center. The registrar validator completed MTQIP orientation and AAAM AIS training. She is employed as a trauma registrar at an ACS-COT verified Level II trauma center. Four total staff have held the MTQIP registrar validator position. Interrater reliability testing ensured consistency across these transitions. The registrar validator does not perform the yearly data validation visit at her trauma center, an alternate registrar from the other three qualified staff is utilized.

Data Validation

The data validation process consisted of the MTQIP program manager and registrar validator reabstracting submitted cases on-site or remotely via temporary EMR access. Data discrepancies were identified and classified according to the following error rating system[9]:

- Type A: Error indicates the validator identified a variable response, but the center did not (e.g., max AIS chest injury severity, validator = 3, center = 0)
- Type B: Error indicates the validator and center both identified the variable response, but disagreed with the answer (e.g., max AIS chest injury severity, validator = 3, center = 2)
- Type C: Error indicates the center identified a variables response, but the validator was unable to confirm medical record documentation consistent with the definition (e.g., max AIS abdomen injury severity, validator = 0, center = 2)

Before reabstraction, the center provided a brief tour of their EMR and data source hierarchy. The MTQIP staff reabstracted 7 to 10 cases over 1 day to 2 days. Validation findings were reviewed with the center staff at the end of the visit. The trauma center was given an opportunity to review and appeal any inadvertent disagreements. In the event of a data element disagreement, the center and validation staff reviewed the data definition. Discrepancies regarding injury severity coding were escalated to the staff at Association for the AAAM for consultation. Issues regarding noncoding variables were referred to the MTQIP program director for final determination. Errors that could be prevented by edit checks remained errors and were routed to vendor staff for software improvement. Circumstances that fell outside of a data definition were clarified in subsequent data dictionary iterations. After a selected case has undergone the entire interrater reliability audit process described, it was considered validated.

A validation summary report was provided to the trauma center and MTQIP executive staff. The summary report included error rates by variable, category, error type, and overall. Recommendations on opportunities for improvement related to noted systematic or documentation issues were included in the report. The trauma center's overall error rate was also a determinant of scoring on the CQI hospital performance index, an annual report card of the center's performance and participation within MTQIP.[11]

Statistical Analysis

Statistical analyses were performed using Stata 14.2 (StataCorp, College Station, TX). Average values were expressed as the mean ± standard deviation. Statistical significance was defined as a p-value of less than 0.05. Cohen's Kappa estimates were calculated for binary (yes/no) variables to assess interrater reliability.[13,14] A sensitivity analysis was performed by repeating the main analysis while excluding high-outlying (i.e., the center's overall error rate was significantly higher than the collaborative-wide error rate) centers from calculations.

This study was submitted to the University of Michigan Medical School Institutional Review Board and given a determination of "not regulated" status as a quality assurance and quality improvement clinical activity.

RESULTS

Study Population

From 2008 to 2018, 185,120 patient submissions met the MTQIP inclusion criteria. Data validation visits were conducted from March 2010 through December 2018. Figure 1 displays the distribution and frequency of cases by visit number and year. Of the 1,243 validated cases, 859 were patients with trauma 50 years or older (Table 2). Clinically, most validated patients sustained

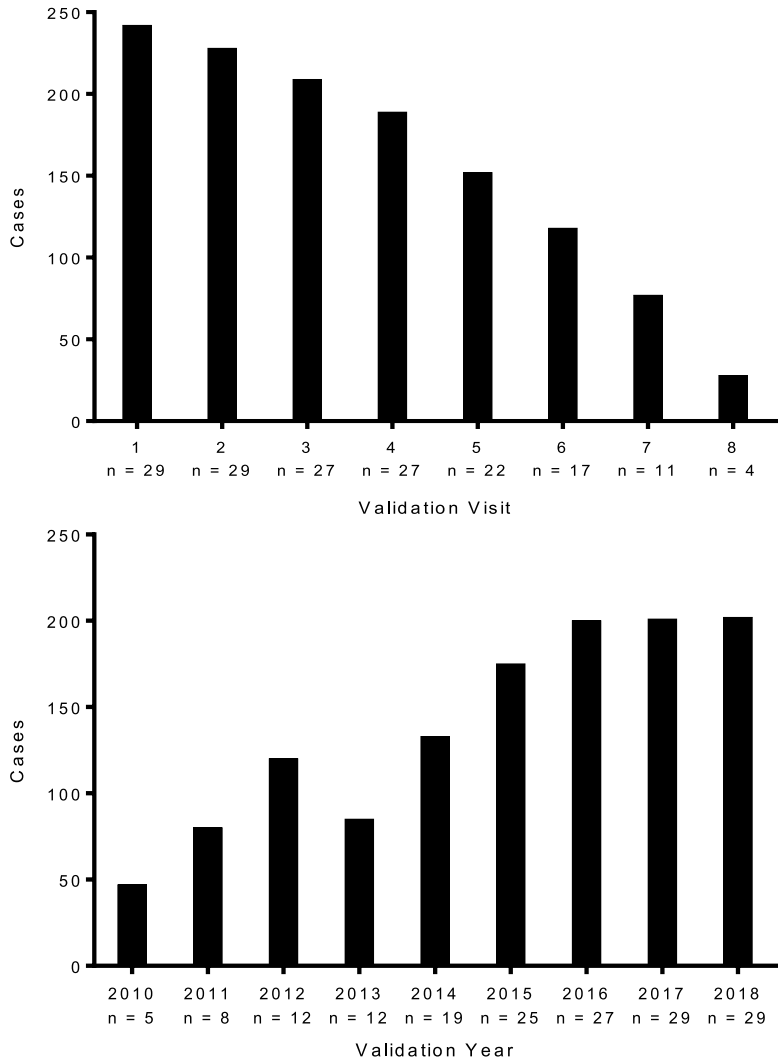

Figure 1. Case volume distribution by (*A*) visit number and (*B*) validation visit year (n = trauma centers).

an injury in the ISS of 5 to 15 range as the result of a blunt mechanism, and required admission to the ICU but did not require mechanical ventilation.

Accuracy

The first validation visit had the highest overall error rate (Table 3). All subsequent validation visits demonstrated a statistically significant decrease in error rate compared with the first visit ($p < 0.05$) (Fig. 2). Errors of omission (type A) were the predominant error type, constituting 2.3% ± 2.8% of errors across all visits. The overall error rate across all validated variables (n = 127,238) was 4.4% ± 3.9%.

In regard to the modeling variables utilized in MTQIP and ACS TQIP benchmark reporting, the data validation program was associated with a statistically significant decrease in the mean error rate for 8 (24%) of the 34 MTQIP and ACS TQIP potential model variables (Table 4). The variables of maximum head/neck AIS severity, hypertension, and functionally dependent health status had the highest overall error rates. The variables of currently receiving chemotherapy for cancer, age, myocardial infarction, and chronic renal failure had the lowest overall error rates. A complete assessment of all 163 variables that were assessed as part of the MTQIP data validation process is provided in Supplemental Digital Content, Table 1, http://links.lww.com/TA/B543.

In a sensitivity analysis, results were recalculated, excluding three centers that had overall error rates higher than the collaborative average. Results were not substantively different from the main (all centers) analysis, demonstrating that trends in error rates were robust to the influence of high-outlying centers.

Reliability

Cohen's kappa estimates were calculated for dichotomous variables to assess reliability. This estimate evaluates reliability by accounting for an agreement that may occur by chance. Estimated kappa reliability of substantial or higher was demonstrated in 90% of the 20 comorbid conditions used in the MTQIP and ACS TQIP modeling. Of the substantial or higher reliability variables, 39% can be attributed to the impact of validation given the statistically significant ($p < 0.05$) interval decrease in error rate

TABLE 2. Patient Demographic and Clinical Characteristics

Category	Validated Cases, n (%)
Sex	
Male	755 (60.7)
Age, y	
< 20	52 (4.2)
20–29	137 (11.0)
30–39	89 (7.2)
40–49	106 (8.5)
50–59	158 (12.7)
60–69	201 (16.2)
≥ 70	500 (40.2)
Race	
Asian	8 (0.6)
Black	196 (15.8)
Hispanic	10 (0.8)
American Indian	3 (0.2)
Other	32 (2.6)
White	994 (80.0)
Missing	0 (0.0)
Mechanism of injury	
Blunt	1,139 (91.6)
Initial ED/hospital pulse rate, bpm	
> 100	341 (27.4)
60–100	788 (63.4)
< 60	84 (6.8)
Missing	30 (2.4)
Initial ED/hospital systolic blood pressure, mm Hg	
≥ 90	1,138 (91.6)
< 90	71 (5.7)
Missing	34 (2.7)
Initial ED/hospital Glasgow Coma Score—total	
14–15	758 (61.0)
9–13	98 (7.9)
3–8	311 (25.0)
Missing	76 (6.1)
ISS	
5–15	637 (51.2)
16–24	189 (15.2)
25–35	339 (27.3)
> 35	78 (6.3)
Mechanical ventilator utilization, d	
0	659 (53.0)
1–7	384 (30.9)
8–14	126 (10.1)
> 14	74 (6.0)
Intensive care unit utilization, d	
0	401 (34.8)
1–7	453 (39.3)
8–14	182 (15.8)
> 14	117 (10.1)
Hospital length of stay, d	
0–7	710 (57.1)
8–14	229 (18.4)
> 14	304 (24.5)

from visit 1 compared with visits 2 through 8. All kappa values in Table 4 were statistically significant ($p < 0.05$).

Of the 37 total validated comorbidities, 78% of variables revealed an estimated kappa reliability of substantial or higher. Fifty-three percent of the 30 validated complications demonstrated an estimated kappa reliability of substantial or higher. Complications with reliability estimates of substantial or higher and the lowest overall error rates were graft/prosthesis/flap failure, wound disruption, abdominal compartment syndrome, deep surgical site infection, pulmonary embolism, extremity compartment syndrome, and enterocutaneous fistula/gastrointestinal leak. A complete assessment of reliability for all validated comorbidities and outcomes is available in SDC 1. All kappa values in SDC 1 were statistically significant ($p < 0.05$) except diabetes mellitus requiring oral therapy ($p = 0.6$), seizure disorder ($p = 0.5$), esophageal varices ($p = 0.5$), atrial fibrillation ($p = 0.5$), and the complication cardiopulmonary arrest ($p = 0.6$).

DISCUSSION

The objective of this study was to assess the impact of a standardized data validation program on data accuracy and reliability over time. The overall error rate demonstrated a statistically significant interval improvement in accuracy throughout all subsequent data validation visits when compared with the index value (6.2% to 4.0%). Errors of under capture (type A) were the most frequent, followed by inconsistent capture (type B) and over capture (type C), respectively. All three types of error demonstrated significant reductions when compared with the initial data validation visit values over time. The majority of dichotomous comorbid and outcome variables validated were found to be of substantial or higher reliability.

Failure to assure adherence to standardized data definitions when collecting clinical outcomes data can lead to loss of program integrity and result in the participants questioning the report findings. The NSQIP acknowledged this fact when the Veterans Health Administration was administering it and considered external data validation through interrater reliability audits to be a central pillar of the program.[10,15] Unfortunately, this commitment to robust data validation utilizing site visits and data reabstraction appears to be waning within the quality programs offered by the American College of Surgeons. The ACS NSQIP has not published results or descriptions detailing its external data validation program since 2010 and now relies on random audits. The ACS TQIP has never published information about its external data validation experience and is now focused on the use of computer audits alone to assess data validity.[16,17]

Accuracy is the degree to which the recorded data correctly reflects the true patient care delivery and disease state. Achievement of accuracy requires ongoing review to ensure cogent entry of local registry data while reducing errors of omission and misinterpretation of information. Serious inconsistencies have been found to exist in current trauma registry data entry.[18–21] This variability has the potential to affect the credibility of data utilized for benchmark reporting and performance improvement. A 15-month study, performed in Georgia by ACS TQIP participants, revealed data heterogeneity was reduced with a program of standardized audit filters and chart review.[22] Across other medical disciplines, standardized

TABLE 3. Mean Hospital Error Rate by Validation Visit and Error Type

Validation Visit	Error Type A*	Error Type B**	Error Type C†	All Error Types	*p*-Value‡
1	3.5 ± 3.6	2.1 ± 2.1	0.6 ± 0.9	6.2 ± 4.7	Reference
2	2.2 ± 2.4	1.7 ± 1.8	0.6 ± 1.1	4.5 ± 3.8	<0.001
3	2.1 ± 2.1	1.4 ± 1.6	0.5 ± 0.9	3.9 ± 3.0	<0.001
4	2.3 ± 3.2	1.3 ± 1.6	0.6 ± 1.1	4.1 ± 3.9	<0.001
5	1.7 ± 2.3	1.5 ± 1.8	0.6 ± 1.3	3.9 ± 3.4	<0.001
6	2.0 ± 2.3	1.3 ± 1.4	0.4 ± 1.0	3.7 ± 3.1	<0.001
7	1.5 ± 2.4	1.2 ± 1.6	0.4 ± 0.8	3.1 ± 3.5	<0.001
8	2.6 ± 3.9	0.8 ± 0.8	0.3 ± 0.5	3.6 ± 4.3	<0.001
All visits	2.3 ± 2.8	1.5 ± 1.8	0.5 ± 1.0	4.4 ± 3.9	

Data represented as % (± SD) unless otherwise noted.
* Error type A indicates the validator identified the variable, but the center did not.
** Error type B indicates the validator and center identified the variable, but disagreed with the answer.
† Error type C indicates the center identified the variable, but the validator was unable to confirm documentation consistent with the definition.
‡ Comparisons were performed for all error types by validation visit (visit 1 vs. subsequent visits 2–8).

interrater reliability audits have been adopted to assess and verify adherence to patient case selection and data definitions within comparative effectiveness programs reliant upon data abstracted from the medical record.[9,10,15,23,24]

For the data elements used as covariates in the MTQIP risk-adjustment modeling, most comorbidity errors improved overall and demonstrated substantial kappa reliability or higher. A previous publication references a 5% error rate as

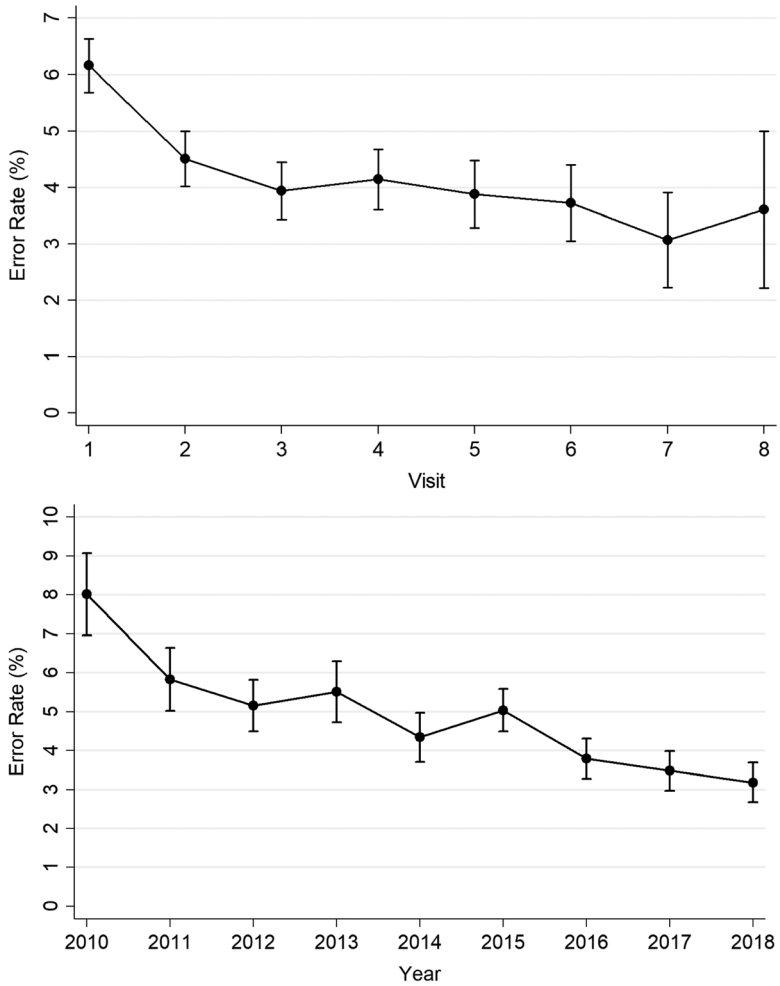

Figure 2. Mean error rate by (*A*) validation visit (n = 166 visits) and (*B*) validation year (n = 166 visits). Error bars represent 95% confidence interval.

TABLE 4. Mean Hospital Error Rate (%) by Validation Visit for Variables Considered for Risk-Adjusted Modeling*

Variables	Visit 1	Visit 2–8	All Visits	*p*-Value**	Kappa†, ‡
Total audited cases (n)	242	1001	1243		
Demographic information					
Age	0.7	0.0	0.4	0.37	
Sex	0.7	0.9	0.8	0.88	
Race	2.2	0.0	1.2	0.12	
Injury information					
Primary external cause code	8.0	4.5	6.5	0.27	
Prehospital CPR	0.7	1.8	1.2	0.44	
ED information					
Initial ED/hospital pulse rate	5.0	5.7	5.6	0.65	
Initial ED/hospital systolic blood pressure	5.8	5.3	5.4	0.76	
Initial ED/hospital GCS—motor	5.0	7.5	7.0	0.17	
Transfer status (direct admit and arrive from)	0.0	1.8	0.8	0.11	
Injury information					
Maximum head/neck AIS severity	14.5	15.0	14.9	0.84	
Maximum face AIS severity	5.8	4.0	4.3	0.22	
Maximum chest AIS severity	8.3	5.8	6.3	0.16	
Maximum abdomen AIS severity	4.1	4.8	4.7	0.66	
Maximum extremity AIS severity	5.0	4.3	4.4	0.65	
Comorbid information					
Alcohol use disorder	7.0	3.9	4.5	0.04	0.71
Current smoker	10.5	6.0	6.5	0.08	0.82
Substance abuse disorder	2.1	4.3	4.1	0.28	0.82
Functionally dependent health status	9.5	8.9	9.0	0.83	0.63
Chronic obstructive pulmonary disease	5.8	3.0	3.5	0.04	0.80
Cirrhosis	2.1	0.5	0.8	0.01	0.70
Congestive heart failure	2.5	1.9	2.0	0.56	0.57
Angina pectoris	1.0	0.4	0.5	0.48	0.76
Myocardial infarction	1.2	0.2	0.4	0.02	0.61
Peripheral arterial disease	0.0	0.7	0.6	0.40	0.66
Hypertension	14.5	7.8	9.1	0.00	0.80
Chronic renal failure	0.4	0.4	0.4	0.98	0.85
Cerebrovascular accident	2.5	1.3	1.5	0.18	0.71
Dementia	3.1	2.3	2.4	0.61	0.84
Mental/personality disorders	14.4	6.7	7.5	0.01	0.76
Disseminated cancer	1.7	0.3	0.6	0.01	0.46
Steroid use	1.7	1.6	1.6	0.95	0.66
Bleeding disorder	6.2	3.6	4.1	0.07	0.69
Currently receiving chemotherapy for cancer	0.4	0.0	0.1	0.04	0.80
Diabetes mellitus	2.5	2.4	2.4	0.94	0.89

CPR, cardiopulmonary resuscitation; GCS, Glasgow Coma Scale. Data represented as % unless otherwise noted. The current variable name is used over historical nomenclature where applicable. Empty cells below kappa indicate a nonbinary variable.

* American College of Surgeons. *TQIP benchmark report references.* Chicago, IL 2016.

** Comparisons were performed for all error types by validation visit (visit 1 vs. subsequent visits 2–8).

† A kappa value of 0.0 to 0.2 indicates slight agreement, 0.21 to 0.4 indicates fair agreement, 0.41 to 0.6 indicates moderate agreement, 0.61 to 0.8 indicates substantial agreement, 0.81 to 1 indicates perfect agreement, and negative values indicate degrees of disagreement.

‡ Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159–174.

a potentially troublesome threshold.[9] While a percentage assessment may be valuable for aggregate trending or evaluation, at the variable level, we do not see that this empiric target correlates with reliability.

Nearly all people are familiar with the famous cinematic moment when Toto pulls back the curtain to expose the blustery man behind the Wizard of Oz. He famously says, "Pay no attention to that man behind the curtain!" Because attention to the quality of the data utilized in a study and how it is used is critical, the editors of JAMA Surgery have published a checklist to elevate the science of surgical database research.[25] Two statements in this checklist are relevant to our study and data validation: 1) ensure that the data variables have not changed over time. If so, account for this. And (2) ensure that data issues, such as missing data, are discussed and that any sensitivity analyses or imputations performed are reported in a clear and cohesive

way. Interrater reliability audits assure not only that data variables have not changed over time but also that the interpretation of the data variable and definition has been applied consistently in multiple hospitals, over time. Holding participants accountable for data capture and entry through reabstraction and data validation auditing makes certain that missing data are kept to a minimum.

Future endeavors should focus on dovetailing the data collection infrastructure with the EMR to directly import variables and flag clinically complex variables. This approach holds the potential to reduce redundant entry, optimize resources, and expand variable capture. It would be valuable to determine which of the audit filters yield the greatest discrepancies and then apply regression modeling to predict challenging cases and provide personalized feedback. An evolving and expanding cadre of abstraction staff should be expected, and systems need to be in place to account for these changes to flatten the learning curve and maintain the fidelity of captured data. Other disciplines have realized gains via a skills-based approach utilizing deliberate practice with immediate feedback or coaching.[26–31]

The inclusion of a data validation program is fundamental to any benchmark reporting effort and has been pivotal to the MTQIP collaborative quality gains.[32] When confronted with results that suggest there is a problem, the natural first response from a trauma center is skepticism. This skepticism is primarily centered around being "different." A robust data validation program assures credibility and removes the ability to remain skeptical based on data validity concerns. Hence, a trauma center can then focus on the quality of care activities to improve their individual outcomes.

This study has several limitations that must be acknowledged. It was confined to participants within Michigan and the collaborative, which may not reflect the education and resources available to other locales. The audit process, while robust, was limited to a select number of patients among the total submitted by each trauma center. The selection criteria utilized, while tailored to provide information on specific high yield situations, could induce bias. Lastly, the infrastructure available to MTQIP facilitates the conduct of interrater reliability audits with regard to resource availability and encouragement of trauma center participation. Collaborative development is ongoing in other states/regions, which may further confirm the findings of this study. The remote data validation approach described, utilizing technology resources, eliminates the need for time-consuming and costly travel, thereby lowering the barrier to conducting these audits. Remote validation is a mechanism that could be scaled across states or collaboratives to assess national data quality.

CONCLUSION

This study addresses the paucity of literature on conducting data validation through interrater reliability audits. Implementation and conduct of the standardized MTQIP data validation program resulted in a statistically significant improvement in the overall error rate in trauma centers across the collaborative. Improved data reliability both within and between trauma centers improved risk-adjustment model validity and quality improvement reporting feedback.

AUTHORSHIP

J.L.J. and A.H.C.-N. had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. J.L.J., M.R.H., A.H.C.-N., P.C.J. participated in the study conception and design. J.L.J., S.L.D.P., A.H.C.-N., M.R.H. participated in the acquisition, analysis, or interpretation of data. J.L.J., M.R.H., A.H.C.-N. participated in the drafting of the article. J.L.J., M.R.H., A.H.C.-N., P.C.J., J.N.M., S.L.D.P. participated in the critical revision of the article for important intellectual content. J.L.J., A.H.C.-N., M.R.H. participated in the statistical analysis. M.R.H., J.N.M., S.L.D.P. participated in the administrative, technical, or material support. J.L.J., M.R.H. participated in the study supervision.

ACKNOWLEDGMENTS

We would like to recognize the following registrars for their dedication to data integrity by serving as MTQIP validators: Jeri Dihle, Susan Huehl, and Cecilia Roiter. We would like to thank the following vendors and staff for providing user support and creation of edit checks: ArborMetrix, Association for the Advancement of Automotive Medicine, Clinical Data Management, Digital Innovation, Inc., Sue Auerbach, Chris Birkmeyer, Zhaohui Fan, Caroline Israel, Dave Karres, John Kutcher, Alex Leaven, Tony Mignano, Cindy Ragland, Jody Summers, and Rob Tewey. We would like to acknowledge the following MTQIP Members for their commitment to improving patient care and data quality: John Fath, James Wagner, Cara Seguin, Tracey Stockinger, Sharon Morgan, and Gail Colton, of Beaumont Hospital-Dearborn; Allan Lamb, Kathy Franzen, Ramona Dinu, and Heather Payton, of Beaumont Hospital-Trenton; Michael Rebock, Barb Smith, Catherine Levinson, Corinna Azar, and Robin Lebeis, of Beaumont Hospital-Farmington Hills; Randy Janczyk, Michelle Schnedler, Holly Bair, and Shannon Zientek, of Beaumont Hospital-Royal Oak; Tom Rohs, Sally Ossewaarde, Sabrina Luke, and Jodie Vining, of Ascension Borgess Hospital; Scott Davidson, Rita Cox, Patricia Benoit, Krisann Woodley, Tonya King-Stratton, Loretta Farrell, Mary Loney, and Dominique Termaat, of Bronson Methodist Hospital; Sujal Patel, Debbie Falkenberg, Kenda Parker, Kristin Wolfgang, Julie MacDougall, and Deanne Krajkowski, of Covenant HealthCare; Anna Ledgerwood, Maidei Munemo, Lisa Salerno, La Toya Kimbrough, Greta Eagger, and Katherine Dhue, of Detroit Receiving Hospital; Brian Shapiro, Zachary Landers, Jennifer Sunderman, and Raquel Yapchai, of Genesys Regional Medical Center; Jeffrey Johnson, Beth Fasbinder, Andrea Nelson, Cheryl Church, and Velma Cuevas, of Henry Ford Hospital; Scott Barnes, Chris McEachin, Michelle Schwarb, and Rose Morrison, of Henry Ford Macomb Hospital; Leo Mercer, Mike McCann, Michelle Maxson, Gloria Lahoud, Shirley Ulmer, and Amber Dombrowski, of Hurley Medical Center; Nicholas Nunnally, Ashley Brown, Alisha Sholtis, and Erin Veit, of McLaren Lapeer Region; Mandip Atwal, Susan Schafer, Marita Vandenberg, Leslie Frezza, and April Pizzo, of McLaren Macomb; John Ketner, Courtney Berry, Megan Wright, and Carolyn Ivan, of McLaren Oakland; Thomas Veverka, Shari Meredith, Tom Wood, Michelle Abedrabo, Teresa Rollin, and Lori Coppola, of MidMichigan Health; Steven Slikkers, Shamarie Regenold, Tanya Jenkins, Allen Stout, and Jill Jean, of Munson Healthcare; Peter Lopez, Joann Burrington, Rebecca Steele, and Carly Callahan, of Providence Hospital; Marco Hoesel, Gwyneth Navas, Melissa Keller, Lisa Zanardelli, TiJuan Davis, Danielle Finn, and Patricia Danhoff, of Sinai-Grace Hospital; John Kepros, Penny Stevens, Kristen Jorae, Paige Harakas, Christopher Stimson, and Maria Maier, of Sparrow Health System; Gaby Iskander, Amy Koestner, Jennifer Haverkamp, Kathy Crystal, Elizabeth Delrue, Gayle Mack, Kristen Thornton, and Kelly Burns, of Spectrum Health; Wayne Vanderkolk, Sherri Veurink-Balicki, Kristi Diephouse, and Coleen Kelly, of Mercy Health Saint Mary's; Joseph Buck, Karrie Brown, Melissa Cunningham, Melissa Jeffrey, Marie Westfall, and Kathleen Waderlow, of St. John Hospital; Wendy Wahl, Mary-Margaret Brandt, Kathy Kemph, Donna Tommelein, Nancy Hofman, and Rebecca Peterson, of St. Joseph Mercy Ann Arbor; Alicia Kieninger, Carol Spinweber, Michele Hunt, and Ellen Noone-Eustice, of St. Joseph Mercy Oakland; Thomas Oweis, Rick Ricardi, Mikki Favor, Jessica Landry, and Ruth Vernacchia, of St. Mary Mercy Livonia Hospital; Samer Kais, Kerri Chernauckas, Kelly Bourdow, Erin Perdue, and Nancy Walter, of St. Mary's of Michigan; Cindy Wegryn, Chris Wagner, Sara Samborn, and Diane Tuttle-Smith, of Michigan Medicine; Larry Lewis, Tammy Luoma, Jodi McCollum, Sarah Sutter, and Lisa Taylor, of UP Health System Marquette.

Funding: This study was supported by a Blue Cross Blue Shield of Michigan and Blue Care Network Collaborative Quality Initiatives grant and a Michigan Department of Health and Human Services grant to Mark R. Hemmila for administration of the Michigan Trauma Quality Improvement Program.
Role of the Funder/Sponsor: The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the article; and decision to submit the article for publication.

DISCLOSURE

J.L.J., S.L.D.P., J.N.M., A.H.C.-N., and M.R.H. receive salary support from Blue Cross Blue Shield of Michigan and Blue Care Network (a nonprofit mutual company) through their support of the Michigan Trauma Quality Improvement Program. P.C.J. is supported by the National Heart, Lung, and Blood Institute Career Development Program in Emergency Care Research K12-KC068853. There are no other conflicts to disclose.

REFERENCES

1. Cook A, Osler T, Glance L, et al. Comparison of two prognostic models in trauma outcome. *Br J Surg*. 2018;105(5):513–519.
2. Glance LG, Dick AW, Mukamel DB, Meredith W, Osler TM. The effect of preexisting conditions on hospital quality measurement for injured patients. *Ann Surg*. 2010;251(4):728–734.
3. Glance LG, Osler TM, Mukamel DB, Meredith W, Wagner J, Dick AW. TMPM-ICD9: a trauma mortality prediction model based on ICD-9-CM codes. *Ann Surg*. 2009;249(6):1032–1039.
4. Osler T, Glance L, Buzas JS, Mukamel D, Wagner J, Dick A. A trauma mortality prediction model based on the anatomic injury scale. *Ann Surg*. 2008; 247(6):1041–1048.
5. Hashmi ZG, Dimick JB, Efron DT, Haut ER, Schneider EB, Zafar SN, Schwartz D, Cornwell EE 3rd, Haider AH. Reliability adjustment: a necessity for trauma center ranking and benchmarking. *J Trauma Acute Care Surg*. 2013;75(1):166–172.
6. Newgard CD, Fildes JJ, Wu L, et al. Methodology and analytic rationale for the American College of Surgeons trauma quality improvement program. *J Am Coll Surg*. 2013;216(1):147–157.
7. Nathens AB, Cryer HG, Fildes J. The American College of Surgeons trauma quality improvement program. *Surg Clin North Am*. 2012;92(2):441–454, x-xi.
8. *Resources for optimal care of the injured patient*. Chicago, IL: American College of Surgeons, Committee on Trauma; 2014.
9. Shiloach M, Frencher SK Jr., Steeger JE, Rowell KS, Bartzokis K, Tomeh MG, Richards KE, Ko CY, Hall BL. Toward robust information: data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program. *J Am Coll Surg*. 2010; 210(1):6–16.
10. Davis CL, Pierce JR, Henderson W, Spencer CD, Tyler C, Langberg R, Swafford J, Felan GS, Kearns MA, Booker B. Assessment of the reliability of data collected for the Department of Veterans Affairs national surgical quality improvement program. *J Am Coll Surg*. 2007;204(4):550–560.
11. Hemmila MR, Jakubus JL. Trauma quality improvement. *Crit Care Clin*. 2017;33(1):193–212.
12. Calland JF, Nathens AB, Young JS, Neal ML, Goble S, Abelson J, Fildes JJ, Hemmila MR. The effect of dead-on-arrival and emergency department death classification on risk-adjusted performance in the American College of Surgeons Trauma Quality Improvement Program. *J Trauma Acute Care Surg*. 2012;73(5):1086–1091; discussion 1091-2.
13. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37–46.
14. Krippendorff K. *Content analysis : an introduction to its methodology*. Beverly Hills: Sage Publications; 1980. 191 p. p.
15. Henderson WG, Moritz TE, Shroyer AL, Johnson R, Marshall G, Ellis NK, Sethi GK, Grover FL, Hammermeister KE. An analysis of interobserver reliability and representativeness of data from the veterans affairs cooperative study on processes, structures, and outcomes in cardiac surgery. *Med Care*. 1995;33(Suppl 10):OS86–OS101.
16. *Validation Summary Report*. Chicago, Ill: American College of Surgeons, Committee on Trauma; 2019. Available from: http://web4.facs.org/tqipfiles/Validation%20Summary%20Report%20Resource%20Guide%20Final.pdf. Accessed 02-21-2019.
17. *Publications*. Chicago, Ill: American College of Surgeons, Committee on Trauma; 2019. Available from: https://www.facs.org/publications. Accessed 02-21-2019.
18. Newgard CD, Fu R, Lerner EB, et al. Deaths and high-risk trauma patients missed by standard trauma data sources. *J Trauma Acute Care Surg*. 2017; 83(3):427–437.
19. O'Reilly GM, Gabbe B, Moore L, Cameron PA. Classifying, measuring and improving the quality of data in trauma registries: a review of the literature. *Injury*. 2016;47(3):559–567.
20. Porgo TV, Moore L, Tardif PA. Evidence of data quality in trauma registries: a systematic review. *J Trauma Acute Care Surg*. 2016;80(4):648–658.
21. Hlaing T, Hollister L, Aaland M. Trauma registry data validation: essential for quality trauma care. *J Trauma*. 2006;61(6):1400–1407.
22. Dente CJ, Ashley DW, Dunne JR, et al. Heterogeneity in trauma registry data quality: implications for regional and national performance improvement in trauma. *J Am Coll Surg*. 2016;222(3):288–295.
23. Eheman CR, Leadbetter S, Benard VB, Blythe Ryerson A, Royalty JE, Blackman D, Pollack LA, Adams PW, Babcock F. National Breast and Cervical Cancer Early Detection Program data validation project. *Cancer*. 2014; 120(Suppl 16):2597–2603.
24. Xian Y, Fonarow GC, Reeves MJ, et al. Data quality in the American Heart Association Get With The Guidelines-Stroke (GWTG-Stroke): results from a national data validation audit. *Am Heart J*. 2012;163(3):392–8, 398.e1.
25. Haider AH, Bilimoria KY, Kibbe MR. A checklist to elevate the science of surgical database research. *JAMA Surg*. 2018;153(6):505–507.
26. Greenberg CC, Dombrowski J, Dimick JB. Video-based surgical coaching: an emerging approach to performance improvement. *JAMA Surg*. 2016; 151(3):282–283.
27. Greenberg CC, Klingensmith ME. The continuum of coaching: opportunities for surgical improvement at all levels. *Ann Surg*. 2015;262(2):217–219.
28. Greenberg CC, Ghousseini HN, Pavuluri Quamme SR, Beasley HL, Wiegmann DA. Surgical coaching for individual performance improvement. *Ann Surg*. 2015;261(1):32–34.
29. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ, Michigan Bariatric Surgery Collaborative. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. 2013;369(15):1434–1442.
30. Bonrath EM, Dedy NJ, Gordon LE, Grantcharov TP. Comprehensive surgical coaching enhances surgical skill in the operating room: a randomized controlled trial. *Ann Surg*. 2015;262(2):205–212.
31. Miglioretti DL, Gard CC, Carney PA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. *Radiology*. 2009;253(3):632–640.
32. Hemmila MR, Cain-Nielsen AH, Jakubus JL, Mikhail JN, Dimick JB. Association of Hospital Participation in a regional trauma quality improvement collaborative with patient outcomes. *JAMA Surg*. 2018;153(8): 747–756.

