Comparing Institutional Trauma Survival to a Standard: Current Limitations and Suggested Alternatives

David E. Clark, MD, MPH

From the Department of Surgery, Maine Medical Center, Portland, Maine, and the Harvard Injury Control Research Center, Boston, Massachusetts.


Trauma centers and trauma systems are naturally interested in measuring their effectiveness, particularly their ability to maximize patient survival after serious injury. A useful measurement of this outcome has been difficult to develop, however, because of the extreme diversity of the trauma population with respect to pre-morbid conditions, types of injury, and geography.

Many studies have evaluated trauma care in hospitals and systems using a "Preventable Death Rate" (PDR), consisting of the number of deaths judged preventable by some person or committee on the basis of autopsy or chart review divided by the total number of deaths during the same time period. This quantity may have some value if the denominator is understood to be a proxy for the total number of patients at risk for preventable death. However, interpreting a difference between two measurements of the PDR, even in the same hospital, may be difficult. For example, an increase in the number of unsalvageable patients will decrease the PDR without any change in the quality of care. In addition, the subjective determination of preventability is notably unreproducible from one expert committee to another; this makes the PDR unsuitable for comparison of different institutions.

Some deficiencies of the PDR could be overcome by using objective measures, such as the Abbreviated Injury Score (AIS) and the Injury Severity Score (ISS), to define a group of patients, and measuring the proportion who survive. One such proposal in a pediatric population was to classify as "severe but salvageable" those patients meeting all of the following criteria: at least one AIS score ≥ 4; either AIS (head injury) ≤ 4 or epidural hematoma; and ISS ≤ 58. Unfortunately, there will still be a significant diversity of risks within such a group, especially when all ages are considered.

Although simple proportions would be easy to understand and would allow for straightforward statistical comparisons, this does not seem to be a useful approach for the trauma population. Instead, much of the progress in predicting outcomes for institutions managing trauma has depended upon predicting the probability of individual patient survival using mathematical models derived from the Major Trauma Outcome Study (MTOS). This approach, which has been called the DEF method (an abbreviation of "definitive"), has been endorsed by the American College of Surgeons. The purpose of this article is to review the DEF method, point out some of its weaknesses, and suggest ways in which the MTOS or other reference databases could be used more effectively to evaluate institutional trauma patient survival.

The DEF Method
The DEF method is based upon the proposal by Flora\textsuperscript{12} that the outcomes of each subject be modeled as a Bernoulli random variable $X_i$ which takes the value 1 if the subject survives (with probability equal to $p_i$) or the value 0 if the subject does not survive (with probability equal to $1 - p_i$). It is not difficult to show mathematically\textsuperscript{13} that if $X_i$ has this distribution, its mean will be $p_i$ and its variance will be $(p_i)(1 - p_i)$. Assuming independence and invoking the Central Limit Theorem,\textsuperscript{13,14} the sum of a sufficiently large number of observed $x_i$ will have approximately a normal distribution with mean $\sum p_i$ and variance $\sum (p_i)(1 - p_i)$, and therefore (Equation 1) will have approximately a normal distribution with mean equal to 0 and variance equal to 1. If it is assumed that additional variation results only from differences between hospitals, this "Z statistic" can theoretically be used for hypothesis testing regarding the quality of care at any given institution. Specifically, absolute values $> 1.96$ allow rejection (with 5\% risk of error) of the null hypothesis that the hospital outcomes are the same as those expected from the given probabilities of survival.

\begin{equation}
Z = \frac{\sum x_i - \sum p_i}{\sqrt{\sum (p_i)(1 - p_i)}} \quad \text{Equation 1}
\end{equation}

A major problem with statistics based on this model is that the true value of $p_i$ for each patient is not known, and valid estimation for individual cases is difficult. Much effort has gone into the attempt to define the probability of survival for any injured patient using multivariate logistic regression models derived from MTOS data. For most of the past decade, the standard regression model has been the Trauma and Injury Severity Score (TRISS), which provides separate equations for penetrating and blunt trauma using age, ISS, and Revised Trauma Score as covariates.\textsuperscript{15}

A more recent multivariate model called A Severity Characterization Of Trauma (ASCOT) introduced additional complexities to the anatomic scoring, and excluded certain groups of patients from the model.\textsuperscript{16} This method has been presented as an improvement on TRISS, in part because of a lower Hosmer-Lemeshow (HL) statistic, which tests whether a logistic regression model fits a categorization of the data from which the model was constructed.\textsuperscript{17} When 10 categories are used, the HL statistic has approximately a $\chi^2$ distribution with 8 degrees of freedom; thus a value $> 15.1$ allows rejection (with 5\% risk of error) of the null hypothesis that the model fits the data. Using a subset of MTOS data, ASCOT (after the exclusions) produced a HL statistic of 24.8 for blunt trauma and 12.65 for penetrating trauma, while TRISS produced HL statistics of 43.9 and 47.5.\textsuperscript{16}

The use of TRISS and ASCOT have been analyzed using other databases;\textsuperscript{18,19} this has shown weaknesses in both models. Hannan,\textsuperscript{19} in particular, discusses problems in the interpretation of the HL statistic. An "independent evaluation" of ASCOT conducted by its originators using a separate database\textsuperscript{20} cited HL statistics of 13.3 for blunt trauma and 20.3 for penetrating trauma as evidence that ASCOT should replace TRISS. Although the authors insisted that a "high degree of agreement is apparent" by graphical comparison of predicted and observed mortality for penetrating trauma, some of this apparent agreement resulted from presenting the data on a logarithmic scale.

Logistic regression models such as TRISS and ASCOT are potentially useful because of the general shape of the logistic distribution and its convenient mathematical properties; however, additional
assumptions are required when the variables in the equation are not binary, or at least carefully
categorized. The fact that a goodness-of-fit test, such as the HL statistic, fails to reject a model does
not ensure that the model's predictions will be accurate over the entire range of probabilities. Many
types of mathematical models can be constructed to predict the probability of survival for the great
majority of cases in which this probability is very high or very low; however, accurate prediction for
individual cases between these extremes will remain difficult even with more advanced modeling
techniques.

A potential problem with these models is that the estimate of \( p_i \) does not necessarily approach the
ttrue value of \( p_i \) as the sample size increases (that is, the estimator may be biased). If this is the case,
then the "Z statistic" based upon this estimate will not approach a standard normal distribution.
Logistic regression estimates of individual probabilities will generally be biased, and will have
more bias when the models (such as TRISS or ASCOT) do not fit the data very well. The actual
distribution of statistics based upon biased estimates of \( p_i \) may be difficult to determine.

In addition to the problem of inaccuracy or bias in the estimation of \( p_i \), there is always the problem of
imprecision, which is only partially resolved by using a very large reference database for construction
of a predictive model. Because of missing data and measurement errors, each \( p_i \) itself may be
considered to have a variance. Although parametric models that do incorporate this additional
variability are mathematically interesting, they add additional complexity to the computations.

Because the "Z statistic" depends, in part, on sample size, it cannot be used to measure the magnitude
of any deviation from the expected number of survivors. MTOS introduced a "W statistic" for this
purpose, defined as (Equation 2) where the predicted number of survivors is \( \Sigma p_i \), the sum of \( p_i \) for
all patients in the sample. The "W statistic" may be interpreted as the number of unexpected
survivors (or if negative, unexpected deaths) for each 100 patients admitted. The "W statistic" is
given a value of zero if \( z \leq 1.96 \); that is, if survival at the index hospital does not differ significantly
from the reference population, it is considered inappropriate to calculate the magnitude of a survival
difference.

\[
W = \frac{\text{actual number of survivors} - \text{predicted number of survivors}}{\text{number of patients} \times 100}
\]

Equation 2

Hollis et al. have pointed out the dependence of the "W statistic" on case mix in each hospital, and
proposed stratification of this statistic in a manner similar to the direct standardization that will be
described in the next section. These authors also demonstrated the value of presenting a summary
measurement along with its confidence interval, using the variance of \( \Sigma x_i \), as an estimate for the
variance of the "W statistic." These definite improvements still have the drawback of using a
potentially biased model to predict the probability of survival for individual cases. Additionally, the
variance of the "W statistic" can only be reasonably estimated by the variance of \( \Sigma x_i \) when the "W
statistic" is close to 0.

The DEF method uses an "M statistic" as a measure of case mix similarity between MTOS and a
given hospital. Both Hollis and Jones present reasons why this statistic is not very useful, and
recommend a standard $\chi^2$ goodness-of-fit statistic or the standardization method described above. Each of these authors also had to base their analyses on the assumption that Boyd's 1987 data on the frequency of MTOS probabilities of survival is still correct,15 because no other description of this distribution has been published.

In summary, although the DEF approach was an important advance in trauma outcome prediction, it cannot be regarded as definitive. Other mathematical models should be explored, although models can only account for a fraction of the variability in the probability of survival for a given case. The wealth of information contained in a large reference database cannot be reduced to a few coefficients in an equation.

**Methods for Stratified Analysis of Cohort Studies**

Standard methods have been developed from many areas of epidemiology for the analysis of cohort studies, in which two groups of subjects are followed in order to observe and compare their outcomes. Usually, as with trauma patients, the outcome variable of interest is binary (e.g., lived or died). Although logistic regression is certainly a valuable tool in this situation, other methods are available which may also be useful when applied to trauma outcomes.

Rather than make mathematical assumptions about the shape of the function relating the probability survival for a given patient to multiple risk factors, we might instead estimate $p_i$ by calculating from observed data (Equation 3) where the carat or "hat" over the $p$ indicates that this is an estimate, and the word "similar" implies that one would have to group or stratify subjects according to the most important risk factors or confounders (e.g., age, severity scores). Using the Bernoulli model and Central Limit Theorem, we could again derive a standard normal test statistic (Equation 4) where $n_i$ denotes the numbers of subjects in each stratum.

$$p_{\hat{i}} = \frac{\text{Survivors (in both groups) with similar risk factors}}{\text{Total patients (in both groups) with similar risk factors}}$$

(Equation 3)

$$z = \frac{\sum x_i - \sum n_i \hat{p}}{\sqrt{\sum n_i (\hat{p} (1 - \hat{p}) / n_i)}}$$

(Equation 4)

By using the actual observed data from a reference population as well as the population of interest (e.g., a hospital), the variability within each stratum of both populations is made part of the model. Furthermore, as $n$ grows larger, these estimates of $p_i$ can be expected to approach the true values of each $p_i$ (that is, the estimators are unbiased). Various mathematical arguments27,28 can be advanced that the true variance of $x_i$ for each stratum will be better approximated if it is multiplied by a "finite sampling factor" (Equation 5) which will be close to 1 if a single hospital is being compared to a much larger reference group. At any rate, if we include this factor and then square the test statistic we get (Equation 6) which is the Mantel-Haenszel test statistic, customarily used for analysis of stratified cohort studies.28,29,30 The square of a standard normal distribution, by definition,13 is a $\chi^2$ distribution with 1 degree of freedom; thus, values greater than 3.84 (the square of 1.96) allow rejection (with 5% risk of error) of the null hypothesis that the test hospital has the same survival rate as the reference population. It should be noted that the Mantel-Haenszel statistic could also be used
for other comparisons; for example, between hospitals or before and after some intervention. This is not possible with the DEF methodology.

The Mantel-Haenszel test, however, does not measure the magnitude of the difference between an index hospital and the reference population. One standard epidemiologic statistic for this purpose is the risk difference (also known as the attributable risk), which would be defined here as (Equation 7) and would range between −1 and 1. If multiplied by 100, this statistic would have the same interpretation as the "W statistic" described earlier. Most epidemiologic studies present the proportion of adverse outcomes (hence the term "risk"), but trauma surgeons have preferred to think in terms of the probability of survival; the same mathematics can be used either way. It is not difficult to show that this estimated difference of two sufficiently large binomial distributions will have an approximately normal distribution with (Equation 8) where \( p_1 \) and \( p_0 \) are estimated as defined above, and \( n_1 \) and \( n_0 \) are the total numbers of subjects in the index hospital and the reference population, respectively.

A crude risk difference is difficult to interpret with heterogenous data. This measurement, however, can also be stratified to enable separate analysis of different strata, and then combined to give a summary measurement of the difference, controlling for the effects of the confounders. This combination assumes that the primary effect of interest (being a patient in the index hospital) is the same in each stratum, and that variations in the estimated risk difference are random. Because the other predictive variables are similar within each stratum, a better estimate of the primary effect can be obtained by pooling the estimates of the primary effect from each stratum. Of course, if the effect is not uniform, a pooled estimate will be meaningless.

Identifying important confounders and properly categorizing them is an art in stratified analysis as well as in regression modeling, requiring more biologic intuition than mathematical sophistication. Analysis of the data presented in stratified form does not require detailed mathematics, and allows the reader to examine whether differences between the index hospital and the reference population are similar in all strata. If the differences (or ratios) are quite variable from one stratum to the next, further investigation of this effect modification may be interesting.

If the effect does seem uniform over all strata, a summary risk difference can be calculated by
weighting the stratum-specific risk differences according to the precision of their estimates and then
adding them, thus \((Equation \ 9)\) where \(F_i\) is the weight for that stratum divided by the total of the
weights for all strata. The generally recommended weight for each stratum\(^{29,30}\) is the inverse of the
variance of the risk difference for that stratum, calculated as above. In the special case where we
might wish to compare different hospitals to a single reference group, we might instead use weights
equal to the fraction of reference group subjects in each stratum; this method of weighting is known
as direct standardization.\(^{29,30}\) Whatever the weighting method used, we can calculate \((Equation \ 10)\)

\[
RD_{\text{summary}} = \sum F_i \cdot RD
\]

\[
\text{Var}(RD_{\text{summary}}) = \sum F_i \cdot \text{Var}(RD)
\]

The methods of stratified analysis presented above are appropriate for comparison between hospitals
or any other two groups of subjects. In general, it is most useful to present the estimate of the
summary risk difference along with its 95% confidence interval, namely the estimate plus or minus
1.96 times the square root of its variance. If this interval includes zero, we can say that the estimate is
not significantly different from zero, and we cannot reject the null hypothesis that there is no
difference between the groups.

While the risk difference is a widely used measure of effect, and is the easiest to explain
mathematically, it may be difficult to envision unless multiplied by an arbitrary number (such as 100)
and converted to a "difference per 100 cases" (like the "W statistic"). Partly because of this
awkwardness, most cohort studies report the risk ratio (RR), which would be defined here as \((p_1/p_0)\),
and easily understood in terms of percentages; if a confidence interval for the RR includes 1, the two
groups are not significantly different. Unfortunately, the derivation of variances and summary
measurements for the RR is more difficult than for the risk difference,\(^{30}\) and will not be presented
here.

The odds ratio, which would be defined as \((p_1/(1 – p_1))/(p_0/(1 – p_0))\), may be used to approximate the
RR; it is the method of comparison used in logistic regression models, but is otherwise not
particularly useful for cohort studies.

In general, some estimate of effect (including its confidence interval) will be more informative than a
hypothesis test. However, it cannot be assumed that an effect is uniform over all strata of a diverse
population.

**Example**

In order to illustrate the methods described above, we can evaluate outcomes from data in the trauma
registry of the Maine Medical Center (MMC) for 1991-1995. The standard DEF statistics calculated
from the MMC data and MTOS equations are \(M = 0.95, Z = 1.03, W = 0\); from these, we can only
conclude that the hospital outcomes are not much different from what would be expected. Adjusting
for case mix as proposed by Hollis,\(^{26}\) we have a standardized \(W_s = 0.46\) with a 95% confidence
interval of \((-2.23, 3.14)\) and the same interpretation.
Because the MTOS authors did report at least a limited categorization of their data by age and ISS, a stratified analysis can also be carried out (Table 1), including calculation of risk difference and RR (along with their 95% confidence intervals) for each of 10 strata.


The Mantel-Haenszel statistic is 24.54, equivalent to a $z$-score of 4.95 (the square root of 24.54), which by itself would lead us to reject (with less than 1% risk of error) the null hypothesis that the probability of survival for MMC patients is the same as for MTOS patients. If the inverse of the variance is used as a weight for each stratum, we obtain a summary risk difference of .0057 with a 95% confidence interval of (.0024, .0089). Because this interval does not include zero, it is also consistent with a "significantly" better outcome; we could interpret this risk difference to mean that out of every thousand patients, we would have about 5.7 "unexpected survivors" at MMC, and we would be approximately 95% confident that the true number of "unexpected survivors" would be between 2.4 and 8.9.

If the direct standardization method is used to weight each stratum, we obtain a standardized risk difference of .0279 with a 95% confidence interval of (.0208, .0350). Again, this is "significant," and could lead to the interpretation that if the distribution of age and ISS at MMC were the same as in the MTOS population, we would actually have about 28 "unexpected survivors" for every thousand patients or 2.8 "unexpected survivors" for every hundred patients. Because we have used the MTOS distribution as a standard, we could compare our standardized risk difference to that of other hospitals that used the same method. It should be noted that this "standardized" estimate might give disproportionate weight to strata that are large in the reference database but small in the test hospital.

Instead of these summary measures, however, a major strength of stratified analysis is the ability to visualize the actual data. A quick study of Table 1 reveals a risk difference (and RR) much higher in the ISS 40-75 categories; all but one of the other categories show either a modestly positive risk difference (with RR > 1) or no significant difference, and there is actually a negative risk difference (with RR < 1) in the older patients with ISS 1-8. The effect of being a MMC patient instead of a MTOS patient is not consistent; it is modified depending upon the category. Although there are mathematical tests for consistency or homogeneity, effect modification may generally be identified by anyone who understands the clinical situation and sees that the effect is in opposite directions or varies in a clinically significant way for different strata.

The presence of effect modification makes any summary measurement suspicious, or at least difficult to interpret. It should lead to further questions about the data, especially the possibility of residual confounding. Indeed, many of the differences might be explained by consideration of another variable (e.g., the Revised Trauma Score, which the MTOS researchers included in the TRISS model but did not report in stratified form). The point is that with the actual data tabulated, any reader can analyze the results rather than depend upon summary statistics. In this example, most would dismiss a "Z statistic" or any single measure of effect, and would require more data before drawing any conclusion about the quality of care at MMC. This further analysis might in fact produce interesting
and previously unsuspected knowledge about differences between hospitals or about trauma outcomes in general.

**Modifications of Stratification and Modeling**

The TRISS model gives separate regression equations for penetrating trauma and blunt trauma, thus creating two strata prior to regression analysis. ASCOT continues this approach, and also analyzes separately certain groups with very high or very low probabilities of survival. Others have proposed separate strata by mechanism\(^\text{19}\) or by risk.\(^\text{26}\) It appears likely that stratified methods will be useful for prediction of individual outcomes even if improved parametric models can be devised.

The major problem with stratified analysis is that when more than a few categories of a few confounders are needed, the number of subjects in each stratum becomes small even with very large databases. For this reason, Miettinen\(^\text{31}\) has suggested that a multivariate confounder score (MCS) can be constructed using logistic regression or similar methods, and that several strata could be constructed from this score to control simultaneously for multiple confounders. The actual data from the hospital under evaluation and the reference population would then undergo stratified analysis as described in the previous section.

An MCS does not have to fit the data perfectly; it need only provide an approximate ranking of the probability of a given outcome, such as survival. Although Miettinen originally proposed that the score be obtained by regression of the combined data of the populations to be compared, a model based upon the reference population alone may suffice,\(^\text{30}\) particularly if it is much larger than the test population. TRISS, ASCOT, or a similar model could thus function as a MCS and provide suitable categories for subsequent stratified analysis. This is essentially the approach taken by Younge,\(^\text{32}\) who modified the "standardized W statistic" by using the actual data from the British version of MTOS, rather than the model predictions, for each stratum.

Miettinen proposed several forms of the MCS, and these have produced some criticism in the epidemiologic literature.\(^\text{29,33-35}\) Although construction of an MCS combines some of the strengths of modeling with those of stratified analysis, it also inherits some of the weaknesses of each, and makes it difficult to assess effect modification by individual confounders. Hypothesis tests after categorization using an MCS have theoretical flaws, but this approach may at least be useful as a graphical method to compare an effect over several categories of risk.\(^\text{35}\)

Another method related to stratified analysis is matching, where each case in the hospital under evaluation is compared to one or more similar cases in the reference population. Obviously, this requires access to the entire reference data-base and a method for selecting sufficiently similar cases. In essence, matching creates many strata, each containing a single case from the index hospital and one or more controls. When the matching is one-to-one, the Mantel-Haenszel test reduces to a very simple form, sometimes known as McNemar's test. If data from individual cases in the reference population are difficult or expensive to obtain, matching may be an efficient approach; however, if data on the entire reference population are readily available, matching offers no advantage over a suitably stratified analysis as described above.\(^\text{29,30}\)

Additional modeling approaches would also be possible if the reference database itself were made available to each hospital. For example, a hospital could create a logistic regression model with a term equal to 1 if the case was from its own data and 0 if the case came from one of the other MTOS...
hospitals. The resulting coefficient on this indicator variable would allow an estimate of the odds ratio comparing survival at that hospital to survival at other hospitals. Access to the reference data would also allow independent investigators to replicate the modeling process that led to the recommended risk-adjustment equations, and thus help to verify and validate the results.

Of course, if the entire reference database were made available to an individual hospital, stratified analyses could also be performed with any desired categorization, and would at least be a helpful adjunct to the evaluation of mathematical models. Even if release of all the reference data were impossible or impractical for proprietary or confidentiality reasons, it would still be useful for MTOS or similar groups to publish detailed summary data using several standard categorizations so that others could compare their hospital data to actual reference data using stratified methods such as those described in this article.

In summary, whether or not a modeling technique is used to compare two groups of patients, stratified data should be presented in tabular form in addition to summary statistics. Comparison to the actual data in a large reference population is more valid than comparison to a mathematical model derived from the reference population. If a comparison is made using a new mathematical model that is too complex to support with tabulated data, the process by which this model was selected should be described in detail. The result of making the same comparison using any previously described standard models should also be reported. Claims of significance resulting from a new, retrospectively constructed model unsupported by presentation of stratified data should be viewed with skepticism.

Additional Considerations

While the focus of this article has been on mathematical methods to estimate random variability and control for confounding, other considerations are equally important in epidemiologic studies.

Selection bias may occur when certain cases are not equally likely to be in the data from the index hospital or the reference population. For example, one hospital might include hip fractures in the elderly, while another does not; if these patients have an unusually high mortality, the first hospital might falsely appear to be "worse."

Information bias is a related concept, referring to the possibility that data may be missing or unobtainable for certain cases. For example, patients who die shortly after discharge or transfer from a hospital may be recorded as survivors. If such defects occur equally in the index and reference populations, they will simply make it harder to detect a difference; if they occur more often in one group, they may seriously distort the measured difference.

Referring back to our example, we might consider the following potential sources of bias: [nlist]

The majority of seriously injured patients at MMC have been transferred from other hospitals, so only those who could be stabilized for transfer are included; this may not be true for MTOS.

The MMC data come from an entire general hospital, including elderly patients with minor injuries admitted to orthopedic or nonsurgical services, while MTOS may consist mostly of patients admitted to trauma services or specialized hospitals.

Only 19,359 MTOS patients with blunt trauma out of a reported total of 64,736 are included in their
published table, these may not be representative of the entire MTOS group.

These and many other questions could be pursued further, which is not the purpose of the present study. It should be clear, however, that attention to the method and completeness of data collection is at least as important as the statistical analysis. To reduce the chance of selection bias or information bias, inclusion and exclusion criteria should be carefully standardized in any multi-institutional study or registry. Along with any numerical analysis comparing two groups of patients, careful analysis of the data quality and completeness in each group is necessary to decide whether the groups are comparable.

Unfortunately, even with attention to all the concerns addressed in this section and this article, comparing institutional trauma survival or other outcomes to a standard will continue to be a challenge. The increasing interest in measuring outcomes of health care has led to extensive research in this area, and risk adjustment for the purpose of comparing hospitals has been found to be a difficult process, even with populations less diverse than our trauma patients. Further study is clearly needed, with a goal of providing methods of outcome measurement that are valid, reproducible, and easily understood.

As we progress, standard methods and terminology used in other areas of biostatistics and epidemiology should be preferred in the study of injury outcomes. Accomplishments and limitations in risk adjustment for trauma outcomes should be compared with those reported from other areas of health services research.

REFERENCES


32. Younge PA, Coats TJ, Gurney D, Kirk CJC. Interpretation of the Ws statistic: application to an integrated trauma system. *J Trauma*. 1997;43:511-515.


Address for reprints: David E. Clark, MD, 190 Park Avenue, Portland, Maine 04102; Fax: 207-774-0459; E-mail: clarkd@poa.mmc.org.