Appendix A: Inverse Probability of Treatment-Weighted Estimators

This approach defines causal effects within the framework of outcomes that would have been observed had subjects been assigned to each of the several possible treatments of interest. In order to study the causal effect of the timing of surgery on the risk of mortality, we would like to compare the risk of mortality corresponding to the hypothetical situation that each patient is assigned to receive definitive fixation after more than twelve hours $P(Y[1] = 1)$ with the risk corresponding to the hypothetical situation that each patient has definitive fixation within twelve hours $P(Y[0] = 1)$. In the setting of a perfect experiment in which patients could be perfectly randomized to treatment before or after twelve hours after admission, a possible estimate of the causal effect of interest is given by the observed relative risk of mortality, $P(Y = 1|A = 1)/P(Y = 1|A = 0)$. As the relationship between the timing of surgery and subsequent clinical outcomes is confounded by baseline covariates such as age, injury severity, or comorbidities, the patients in our sample who were operatively managed after more than twelve hours may not be representative of the entire population of patients. Therefore, the risk of mortality among these patients, $P(Y = 1|A = 1)$, cannot be used as an estimate of the true risk of mortality $P(Y[1] = 1)$ had every patient been assigned to surgery after more than twelve hours.

Robins[29] introduced a class of estimators that address this problem through a straightforward weighting approach. Like propensity score methods[30,31], these inverse probability of treatment-weighted (IPTW) estimators make use of an estimate of the treatment mechanism. IPTW estimators use the probability that a given subject (i) would have received *his or her* observed treatment (A_i) given his or her baseline covariates

(W_i). They then weight each observation by the inverse of this probability, $P(A_i|W_i)$, so that each subject's weight is:

$$w_i = 1/P(A_i|W_i)$$

This creates a new sample in which treatment assignment is independent of the baseline covariates, making it straightforward to estimate $P(Y[1] = 1)$ and $P(Y[0] = 1)$ by fitting of a saturated, weighted logit model. In the setting of a categorically defined treatment time, t_0 through t_4, the following model was fit with corresponding indicator variables (I_a) in order to estimate the risk of mortality at each later interval versus the baseline (definitive fixation within twelve hours):

$$\text{logit}(P[Y(a) = 1]) = B_0 + B_1*I_{(a = t1)} + B_2*I_{(a = t2)} + B_3*I_{(a = t3)} + B_4*I_{(a = t4)}$$

In order to study the possibility of a different causal effect (effect modification) of treatment time separately for subpopulations defined by a variable such as the presence or absence of a serious chest injury (Abbreviated Injury Score <3 compared with ≥3), $P(Y[1] = 1)/P(Y[0] = 1)$ could be calculated within each level of chest injury severity and compared. Effect modification by baseline factors (V_j) was assessed using the following conditional saturated weighted logit model with treatment by covariate product terms:

$$\text{logit}(P[Y(a) = 1]) = B_0 + B_1*I_{(a = t1)} + B_2*I_{(a = t2)} + B_3*I_{(a = t3)} + B_4*I_{(a = t4)} + B_5*V_j + B_6*I_{(a = t1)}*V_j + B_7*I_{(a = t2)}*V_j + B_8*I_{(a = t3)}*V_j + B_9*I_{(a = t4)}*V_j$$

There are two basic assumptions for valid estimation of causal effect using IPTW. The first is that there are no unmeasured confounding variables. This means that the counterfactual

outcome is conditionally independent of treatment, given measured covariates. To the extent that unknown confounding goes unadjusted for, or is uncorrelated with measured confounders, estimates may be biased. This assumption is ubiquitous in observational studies and cannot be empirically tested. The second assumption has been called the experimental treatment assignment (ETA) assumption and is similar to Rosenbaum and Rubin's "strongly ignorable treatment assignment" assumption for use of the propensity score[30]. It requires that there are no values of the baseline covariates for which treatment is assigned in a deterministic fashion. If, for example, patients with head/neck injury scores of 3 or more are always operated on after more than twelve hours, none of the patients in the re-weighted sample who were operated on within twelve hours will have head/neck injury scores of 3 or more, leading to a biased estimate of the corresponding risk of mortality. In this case, the comparison of interest cannot realistically be made among the population at hand because one group of patients could never realistically have received the treatment that they did not receive based on one (severity of head/neck injury) or more characteristics.

Appendix B: Data-Adaptively Selected Treatment Models

IPTW estimators only give consistent estimates of the parameters of interest if the treatment mechanism itself is estimated consistently. Since a misspecified parametric model for the treatment mechanism will lead to inconsistent estimation of the treatment mechanism and thus inconsistent estimation of the causal parameters of interest, we avoided assuming an a priori functional form and instead employed a data-adaptive model selection algorithm that chooses the functional form on the basis of the information that is available in the data at hand. For this purpose, we used a model-

selection approach that is based on polynomial spline functions and includes testing of all one-way interaction terms between candidate covariates[32]. The consistency of the estimates based on the selected model is not affected by adding additional terms for variables to this model that might in truth not be related to the treatment variable. If these variables are, however, associated with the outcome of interest, including them in the treatment mechanism model will adjust for empirical confounding by any of these variables and thus increase the efficiency of the IPTW estimates. On the basis of this observation, we complemented the selected model with main-effect terms for all baseline covariates in W_m that had not already been selected by the data-adaptive model selection algorithm. The data-adaptively selected treatment models, before addition of main-effect terms for other predictors of the outcome, are given in Tables E-1 and E-2.

Data-adaptive model selection based on polynomial splines was performed using the polyspline package in R version 2.3.1[39,40]; multinomial regression models were fitted using the multinom() function in the nnet package; the Hosmer-Le Cessie test was carried out using the resid() function of the Design package.

TABLE E-1 Summary of the Estimates of the Treatment Mechanism That Were Used for the Categorical Mortality Analysis

Covariate	>12 to 24 hr	>24 to 48 hr	>48 to 120 hr	>120 hr
New Injury Severity Score	1.00	1.01	1.03	1.04
Arrival between 6 A.M. and 12 P.M.	0.23	0.77	0.70	0.94
Cardiac comorbidity	1.34	1.70	2.49	2.14
Number of serious associated extremity/pelvic injuries	0.69	0.80	0.91	0.91
Maximum Abbreviated Injury Score, head/neck region	1.02	1.02	1.03	0.98
Bilateral femoral fracture	0.29	0.77	0.57	0.81
Age	1.01	1.01	1.01	1.01
Teaching hospital	1.53	0.91	0.65	1.07
Glasgow Coma Scale score	1.00	0.96	0.95	0.92
Treated at level-1 trauma center	0.95	1.36	1.42	1.40
Cerebrovascular comorbidity	1.24	2.51	1.04	0.00
Hospitals from Northeast region	1.23	1.15	1.76	0.93

The entries in the first column give the factor by which the relative risk of being assigned to surgery between twelve and twenty-four hours rather than to surgery within twelve hours changes for each unit increase in the covariate under consideration. For covariates with factors greater than 1.00, the relative risk of being assigned to the former treatment category rather than the latter one (reference category) thus increases as the value of the covariate increases. Entries in the remaining columns are interpreted accordingly. The polychotomous logistic regression model selected for the treatment mechanism consisted entirely of main-effect terms with the Hosmer-Le Cessie test showing acceptable model fit for all time frames except for >120 hours (p values of 0.24, 0.56, 0.24, and 0.028, respectively).

TABLE E-2 Summary of the Estimated Treatment Mechanism That Was Used for the Binary Mortality Analysis (Investigation of Effect Modification)

Covariate	Odds Ratio (95% Confidence Interval)*
Arrival between 6 A.M. and 12 P.M.	0.52 (0.43 to 0.63)
New Injury Severity Score	1.03 (1.02 to 1.04)
Cardiac comorbidity	1.69 (1.37 to 2.07)
Number of serious associated extremity injuries	0.77 (0.70 to 0.84)
Age	1.01 (1.01 to 1.02)
Glasgow Coma Scale score	0.97 (0.95 to 0.99)
New Injury Severity Score (knot 50)†	0.95 (0.92 to 0.98)
Maximum Abbreviated Injury Score head/neck region	1.01 (0.96 to 1.07)
Bilateral femoral fracture	0.59 (0.30 to 1.14)
Teaching hospital	1.07 (0.91 to 1.25)
Treated at level-1 trauma center	1.18 (1.01 to 1.39)
Cerebrovascular comorbidity	1.54 (0.33 to 7.24)
Hospitals from Northeast region	1.25 (0.95 to 1.64)

*The entries give the estimated odds ratio for being assigned to surgery after twelve hours as compared with surgery during the reference time frame. †Apart from main-effect terms for the covariates that were found to be predictive of mortality, this model contains a spline function for the New Injury Severity Score with a knot at 50, suggesting that the dependence of treatment assignment probabilities on this variable may differ depending on whether the New Injury Severity Score is above or below 50. The Hosmer-Le Cessie test showed acceptable goodness-of-fit (p = 0.54).

Appendix C: Assessing the Validity of the Experimental Treatment Assignment Assumption

Standardized Risk Ratio Approach

Modified IPTW weights were proposed by Sato and Matsuyama[35] in order to expand causal contrasts afforded by IPTW estimates to specific components of the larger study group. Because subjects managed after twelve hours tended to have slightly higher injury severity (Table I), regardless of the extensive confounding control, unknown or unmeasured factors may still preclude some patients from later treatment groups from receiving early treatment (a practical ETA violation). A corollary of this is that only those patients managed early could have reasonably been

treated in any of the five treatment groups (t_0 to t_4) and may represent the only patients who could ethically be randomized. Standardized risk ratios give the closest approximation to such an experiment by focusing on the patients who were managed early. All components of the analysis described in Appendices A through C were identical for the standardized risk ratio (SRR) analysis, except that the calculation of weights was different in order to provide inferences as to the estimated effect of treatment within the early treatment group. Weights were generated with use of the same treatment mechanism model for the IPTW analysis to estimate the conditional probability of early treatment as the numerator and the conditional probability of receiving the treatment received as the denominator:

$$w_i = P(A = t_0|W = W_i) / P(A = A_i|W = W_i)$$

By using the early treatment group (t_0) as the standard population and the mortality experience of the entire study sample, the standardized risk ratio is interpreted as the estimated risk ratio if those who were actually managed early were to have been managed at one of the later treatment time periods.

Monte Carlo Simulation Approach

Wang et al.[36] proposed the following simulation-based approach for examining the extent to which IPTW estimators might be biased due to a violation of the ETA assumption. First, an estimate of the data-generating distribution is obtained that allows one to simulate realizations of the observed data structure. For this estimated data-generating distribution, the true parameter values can be computed through G-computation. A sampling distribution of IPTW estimates can be obtained by applying the IPTW estimator to a large number of simulated realizations of the observed

data structure. Since the assumption of no unmeasured confounding is trivially satisfied in this case, any discrepancy between the mean of these estimates and the true parameter value reflects a violation of the ETA assumption.

In the case of a point-treatment study, the observed data structure O = (W, A, Y) consists of baseline covariates W, treatment A, and outcome Y. An estimate of the data-generating distribution thus consists of an estimate of the marginal distribution of W, the conditional distribution of A given W, and the conditional distribution of Y given A and W. We estimate the marginal distribution of W by its empirical distribution and use the data-adaptive approach described in Appendix B to obtain an estimate of the conditional distribution of A given W. For the mortality analysis, the conditional distribution of Y given A and W is estimated by means of logistic regression. This model includes main-effect terms for A as well as all baseline covariates that were found to have significant univariate associations with the outcome. These estimates now allow us to simulate realizations of the observed data structures by using the following sequential approach. We first generate n realizations of W by sampling with replacement from the n observed values in our dataset. We next draw n realizations of A from the estimated conditional distribution of A given these simulated values of W. Last, we obtain n realizations of Y by drawing from the estimated distribution of Y given the simulated values of A and W. The true parameter values for this data-generating distribution can be computed based on the following G-computation approach. If we want to estimate the counterfactual mortality for the scenario that every patient undergoes surgery within twelve hours of admission, for example, we first draw a large number, say N = 10,000, realizations of W as above; then we

generate N realizations of Y by drawing from the estimated conditional distribution of Y given A = a* and these simulated values of W, where a* represents the treatment level corresponding to surgery within twelve hours. The desired counterfactual mortality can then be estimated by simply taking the mean of these simulated Y values.

Table E-3 summarizes the results of such a simulation study for determining the extent of bias to which our IPTW estimators of the marginal counterfactual mortality risks might be subject because of a violation of the ETA assumption. Figure E-1 shows the corresponding distributions of IPTW estimates relative to the true parameter value obtained by G-computation. In all cases, the bias due to a possible violation of the ETA assumption appears to be minimal. As all subjects seem to have adequately large probabilities of following any one of the five categories of treatment time that we are examining here, this should also be true for the two treatment categories defined by the binary version of treatment so that the corresponding parameters in the binary analysis can also be estimated without appreciable bias due to a violation of the ETA assumption.

TABLE E-3 Estimate of ETA Bias Based on Monte Carlo Simulation

	≤12 hr	>12 to 24 hr	>24 to 48 hr	>48 to 120 hr	>120 hr
G-computation truth	4.15%	2.27%	3.55%	2.68%	3.03%
Mean IPTW estimate	4.16%	2.20%	3.53%	2.64%	2.92%
Estimated bias	0.01%	–0.07%	–0.01%	–0.03%	–0.11%

Appendix D: Modified Bootstrap Approach to Obtaining Confidence Intervals and P Values

Observations from the same hospital are likely to be correlated with each other. In the presence of such correlated data, standard mean estimates are still consistent even though they may no longer be efficient. Therefore, we do not need to take into account the correlation among patients from the same hospital for the purpose of obtaining point estimates. Confidence intervals provided by these standard methods, however, are no longer reliable in the presence of correlated data; they would be based on the assumption of a sample of independent observations and would thus tend to overestimate the information available in the data, resulting in confidence intervals that are too small. Therefore, we estimated confidence intervals with use of the following modified bootstrap approach[37]. During each bootstrap iteration, we draw a sample of size N with replacement from the pool of N hospitals in our dataset to obtain a bootstrap dataset that consists of all patients from the selected hospitals. This is a slight modification to the standard bootstrap approach for independently sampled observations, which would prescribe us to draw samples of size n with replacement from the pool of n patients in our dataset. We follow the same modified bootstrap approach to obtain p values by applying the general resampling-based methodology developed by Pollard and van der Laan[38].

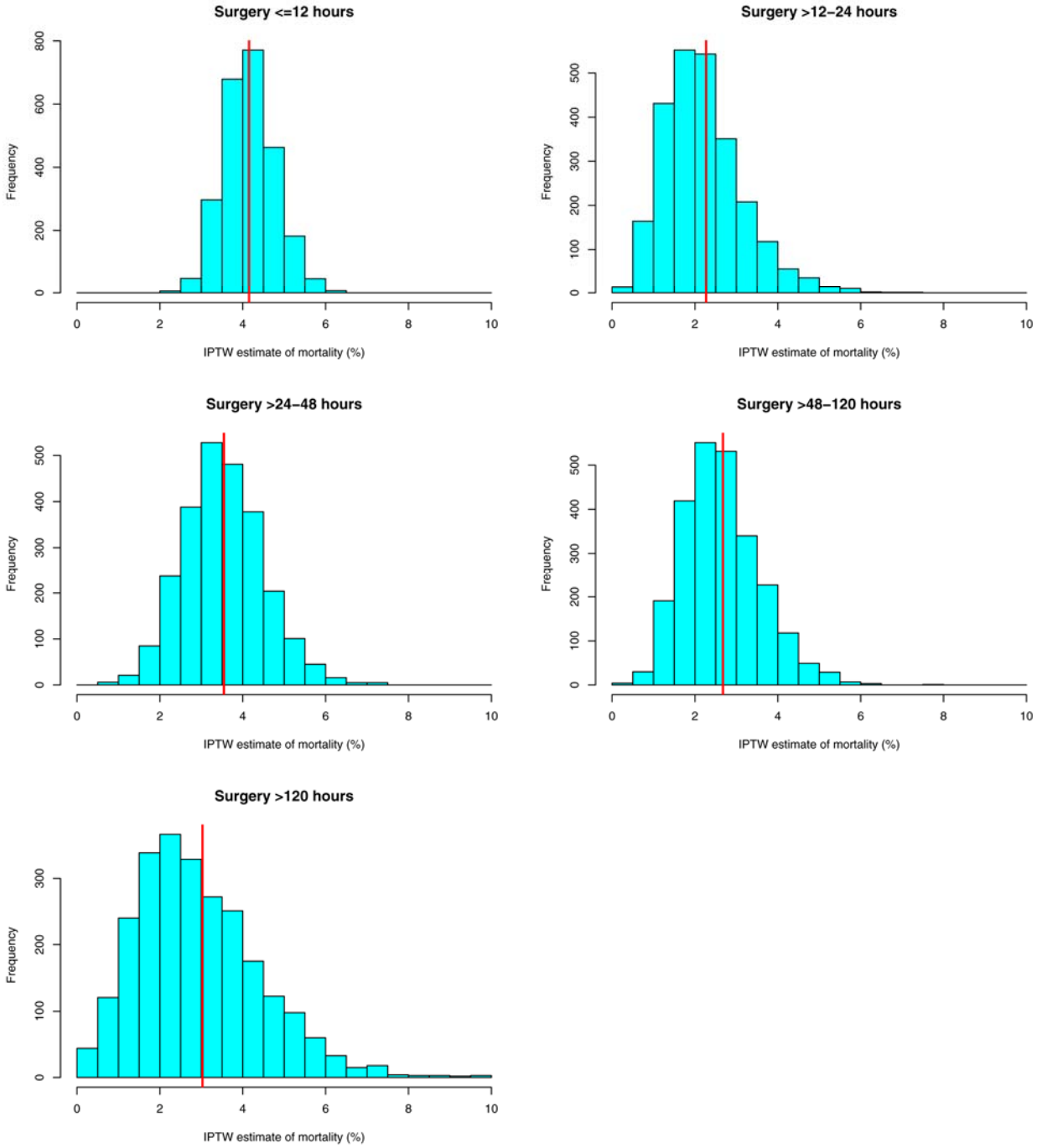

Fig. E-1

The distributions of inverse probability of treatment-weighted (IPTW) estimates relative to the true parameter value obtained by G-computation (vertical red line).