Marginal structural models demonstrate treatment time effect on mortality in multi-system trauma patients with femoral shaft fracture
Marginal structural models demonstrate treatment time effect on mortality in multi-system trauma patients with femoral shaft fracture

by

Saam Morshed

A dissertation submitted in partial satisfaction of the Requirements for the degree of Doctor of Philosophy in Epidemiology in the Graduate Division of the University of California, Berkeley

Committee in charge:
Professor John M. Colford, Chair
Professor Alan Hubbard
Professor Tony M. Keaveny

Spring 2010
Marginal structural models demonstrate treatment time effect on mortality in multi-system trauma patients with femoral shaft fracture

© 2010

by Saam Morshed
# Table of Contents

**CHAPTER 1: OPTIMAL TIMING OF FRACTURE STABILIZATION AMONG MULTISYSTEM TRAUMA PATIENTS: LITERATURE REVIEW**

- Review of Clinical Studies 1
- Methodological Limitations 3
- Study Question 5

**CHAPTER 2: CONVENTIONAL AND COUNTERFACTUAL APPROACHES TO CONFOUNDING AND COVARIATE ADJUSTMENT IN OBSERVATIONAL STUDIES** 6

- Basics Principles of Statistical Inference 6
- Bias, Confounding, and Counterfactuals 7
- Controlling Confounding 8
- Comments 15

**CHAPTER 3: APPLICATION OF MARGINAL STRUCTURAL MODELS TO ESTIMATE CAUSAL EFFECTS OF TIMING OF TREATMENT OF FEMORAL SHAFT FRACTURE** 24

- Study Cohort and Data Source 24
- Treatment Variables 25
- Main Outcome Measure 25
- Statistical Methods 25

**CHAPTER 4: TESTING THE EXPERIMENTAL TREATMENT ASSIGNMENT ASSUMPTION** 31

- Graphical Evaluation of the Experimental Treatment Assignment Assumption 32
- Monte Carlo Simulation Approach 32
- Standardized Risk Ratio Approach 33

**CHAPTER 5: RESULTS** 38

- Inverse Probability of Treatment-Weighted Analysis 38
- Logistic Regression Analysis 39
- Standardized Risk Ratios Analysis 39

**CHAPTER 6: DISCUSSION** 50

- Clinical Comments 50
- Methodological Comments 52
- Final Comments 54

**REFERENCES:** 55
For Nooshin, Kea and Darya who allowed this endeavor, I am grateful.

I love you.
Abstract

Marginal structural models demonstrate treatment time effect on mortality in multi-system trauma patients with femoral shaft fracture

by

Saam Morshed

Doctor of Philosophy in Epidemiology

University of California, Berkeley

Professor John M. Colford, Chair

Context: Fractures of the femoral shaft are common and have potentially serious consequences in patients with multiple traumatic injuries. The appropriate timing of fracture repair is controversial.

Objective: To assess the effect of timing of internal fixation on mortality in multi-system trauma patients

Design: Retrospective cohort.

Setting: Data from public and private Trauma Centers throughout the United States reported to the National Trauma Data Bank (version 5.0 for 2000-2005).

Patients: 3069 multi-system trauma patients (Injury Severity Score $\geq 15$) who underwent internal fixation of femoral shaft fractures.

Intervention: Time to treatment was defined in categories as the time from admission to internal fixation: $t_0 \leq 12$ hours; $12 < t_1 < 24$; $24 < t_2 \leq 48$; $48 < t_3 \leq 120$; $120 < t_4$.

Main Outcome Measure: The relative risk of in-hospital mortality comparing the four later periods to the earliest one was estimated using logistic regression, multivariate standardized risk ratios, and Inverse Probability of Treatment Weighting (IPTW). Subgroups with serious head, chest, abdominal and additional extremity injury were investigated.

Results: The mortality risk estimated by IPTW was significantly lower in several time categories: $t_1$ (RR=0.45, 95%CI: 0.15 – 0.98, $p=0.03$), $t_2$ (RR=0.83, 95%CI: 0.43 – 1.44, $p=0.49$), $t_3$ (RR=0.58, 95%CI: 0.28 – 0.93, $p=0.03$) and $t_4$ (RR=0.43, 95%CI: 0.097 – 0.94, $p=0.03$). Those with serious abdominal trauma (Abbreviated Injury Score, AIS$\geq 3$) experienced the greatest benefit from delay of internal fixation beyond 12 hours (RR=0.82 for AIS$<3$, 95%CI: 0.54 – 1.35; versus RR=0.36 for AIS$\geq 3$, 95%CI: 0.13 – 0.87; $p$-value for interaction $=0.09$). The results of the logistic regression and multivariate standardized risk ratios agreed
with IPTW estimates. Qualitative and quantitative testing of modeling assumptions supported the validity of the marginal IPTW estimates.

**Conclusion:** Delayed repair of femoral shaft fractures beyond 12 hours in multi-system trauma patients significantly reduced mortality for three of the four treatment periods studied. Patients with serious abdominal injury benefit most from delayed treatment. This work applies causal inference methodology and assumption testing that is well suited to orthopaedic therapeutic studies when randomization is not possible or ethical.
Surgical treatment of orthopaedic injuries in severely injured patients has been one of the most significant advances in trauma care in the past century. This achievement has been exemplified by internal fixation of femoral shaft fractures which markedly reduced patient discomfort and the morbidity of protracted skeletal traction and immobilization. While femoral shaft fractures may be surgically stabilized using a variety of methods including external fixation (trans-cutaneous pins linked by bars) and plate osteosynthesis (plate fixation using bone screws), medullary fixation (metal rod inserted in the long-bone medullary canal) has been the most successful method with healing rates approaching 100% with minimal complications. First performed by Kuntcher in 1939, the medullary nail was brought to North America by repatriated prisoners of war and the technique slowly followed with the first clinical series published by Clawson and Hansen. Since that time this technique has become the “gold standard” definitive surgical treatment for fracture of the femoral shaft. The clinical result from observational studies have been so dramatic and reproducible that no randomized controlled trial has ever been undertaken to establish superiority of operative versus non-operative management of femoral shaft fracture. However, the optimal timing of surgical management of long-bone fractures in multi-system trauma patients is controversial. This controversy has persisted because of inconsistency in way that the question has been framed as well as methodological inadequacies in how researchers have attempted to answer it.

Review of Clinical Studies

Several observational studies and one randomized study have suggested clinical benefits from early stabilization of major orthopaedic injuries—those of the shaft of the femur being most common—in reducing the incidence of pulmonary complications and mortality. Riska was one of the first authors to show a reduction in fat embolism syndrome (FES) in multi-system trauma patients undergoing early (time not defined) fixation of long-bone, pelvic and spine fracture versus a historical conservatively (non-operative) treated control group with similar injury characteristics (4.5% versus 22%). The endpoint of this and many other similar studies was FES, an inconsistently defined clinical syndrome thought to be due to the embolization of marrow contents to the lungs manifested by: “snow-storm” infiltration on chest radiograph, arterial partial pressure of oxygen less than 60mmHg, skin petechial hemorrhages, aberrance of consciousness, progressive anemia, tachypnea, and thrombocytopenia. Given that the validity or reliability of this definition has never been proven and that acute lung injury in the post trauma setting is now believed to be secondary to many other factors, references to FES have mostly disappeared from the literature. A more current, but still flawed definition of pulmonary end-organ injury is Adult Respiratory Distress Syndrome (ARDS) which is determined on the basis of: 1) acuity, 2) bilateral infiltrates on chest radiographs without a cardiogenic cause (i.e. no left heart failure), 3) an arterial partial pressure oxygen to fraction of inspired oxygen ratio of less than 200. Again, this definition has been shown to have limited intra-observer reliability but has served to define pulmonary end-organ injury in the majority of studies assessing the importance of timing of fracture fixation among multi-system trauma patients. Johnson, in a retrospective observational study of patients with multi-system trauma...
and at least two major long-bone fractures, found that early operative stabilization was associated with a five fold increase in the odds of ARDS when controlling for the Injury Severity Score.

The Injury Severity Score (ISS)\(^{18, 19}\) is one of the first and most commonly used of many systems of grading overall injury severity and was borne from efforts to improving triaging of patient during the early stages of regionalization of trauma systems in the Unites States. It has since become an instrumental tool in risk adjustment for research on trauma patients. It is comprised of the Abbreviated Injury Score (AIS)\(^{20}\) which assigns a severity grade from minor (0) to unsurvivable (6) to each of a patient’s injuries. The body is divided into six regions—head and neck, face, chest, abdomen, extremity and pelvis, skin—and the highest AIS graded injury from the three most severely involved regions are summed to give the ISS. The ISS has been criticized for it’s non-additivity, non-parametric distribution, and failure to account for more than one severe injury in a given region\(^{21-23}\). While improvements to the ISS have been suggested such as the New Injury Severity Score (NISS)\(^{19}\), which sums the squares of he highest AIS regardless of body region, the use of the ISS is ubiquitous in injury research pertaining to optimal timing of fracture care.

As definitive internal fixation techniques were popularized, enthusiasm for their apparent ability to allow patients who would have otherwise been immobilized for weeks or months to mobilize earlier expanded into a belief that earlier surgery was causally responsible for improving outcomes. Behrman\(^{13}\) performed a retrospective cohort study of multi-system trauma patients with femur fractures and found that those treated early (less than 48 hours from admission) had reduced pulmonary complications included ARDS and pneumonia. Despite the observational nature of the study and lack of any formal confounding adjustment, these authors concluded, “Delayed femur fixation increased the incidence of pulmonary complications.” Similar unfounded causal assertions were made by Charash\(^{14}\) who found few complications and decreased total hospital and intensive care unit length of stay in patients treated less than 24 hours from admission with medullary nails.

Bone\(^{16}\) performed the first and only published randomized controlled trial testing the hypothesis that early definitive management of femoral fractures reduced morbidity. They randomized 178 patients with femoral fractures, 83 of which had an ISS of greater than 18 (one of several thresholds defining the multi-system trauma patient). Despite small sample size, they were able to show a significantly higher rate of the composite endpoint of ARDS, pulmonary dysfunction, or FES in patients randomized to late (more than 48 hours from admission) treatment. However, there was a ten-fold higher rate of chest injury at the time of admission that was not balanced by the randomization or adjusted for in the analysis. These initial chest injuries may have led to higher rates of the pulmonary complications in the late treatment group.

While it has been shown that patients with isolated femoral shaft fracture have reduced morbidity with early treatment\(^{24}\) and perhaps those with multiple injuries and higher ISS do as well as suggested by the aforementioned papers, the question has been posed as to whether certain associated injuries patterns may define subgroups of patients that do poorly with early treatment. Pape\(^{25}\) studied 106 patients with mid-shaft femoral fractures and an ISS greater than 18. Patients were treated either prior to or later than 24 hours from admission and stratified by presence or absence of chest trauma (AIS greater than or equal to 2). Those with chest injury treated early had a higher incidence of ARDS (p<0.05) and mortality (not statistically significant) than those without severe chest injury that were treated early, though a formal test of
interaction was not reported. These authors and others\textsuperscript{26, 27} have found higher incidence of inflammatory deregulation (primarily increased serum interleukin-6) and organ dysfunction (ARDS and pneumonia) with treatment up to five days from admission. Similar susceptibility to ongoing injury from early surgery has been suggested among multi-system trauma patients with severe head injuries (AIS greater than or equal to 2)\textsuperscript{28} though how acute physiologic perturbations translate into long-term functional effects has been questioned\textsuperscript{29}.

Studies that suggest adverse effects of early treatment among the most badly injured patient such as those with associated severe head or chest injury attempt to establish coherence with the “two-hit” model for post traumatic organ failure where systemic hypo-perfusion\textsuperscript{30} and inflammation\textsuperscript{31-33} may increase susceptibility to end-organ injury and multiple organ failure if further stress (iatrogenic in the case of fracture surgery that typically involves large volume blood loss and systemic embolization of marrow contents) is prematurely undertaken. Several different pathways have been implicated in the development of end-organ injury and MOF in multi-system trauma patients, yet no central paradigm of causation has emerged. The degree and duration of hypo-perfusion, or inadequate resuscitation (with fluids, electrolytes, and/or blood products), is known to be an important risk factor in predicting subsequent complications among the critically injured\textsuperscript{34, 35}. Hypo-perfusion is typically measured by global measures of oxygen delivery, namely arterial base deficit or lactate\textsuperscript{36}. Several investigators have found that admission base deficit and lactate correlated with increased need for blood transfusion, length of hospitalization, and end-organ injury such as ALI\textsuperscript{37, 38}. Because the optimal method for the assessment of the adequacy of resuscitation is not clear\textsuperscript{36} it is believed that hypo-perfusion often remains occult and unrecognized prompting increased emphasis on early and aggressive resuscitative efforts prior to major secondary surgery such as internal fixation of long-bones. Still, how long-bone fractures and the care contribute to these complex metabolic processes is still poorly understood.

**Methodological Limitations**

In an attempt to draw inferences to guide clinical decision-making based on extant clinical literature on the topic of optimal timing for femoral fracture fixation in multisystem trauma patients, several systematic reviews have been undertaken. A meta-analysis\textsuperscript{39} found a large relative risk reduction (RR 0.30, 95% confidence interval 0.22 – 0.40) in respiratory complications with early operative fixation (usually within 24 hours). However, quantitatively pooling these studies may not be appropriate. Differences in treatment definition (both time cut-offs and type of fixation used) and outcome (mortality versus length of hospitalization versus unreliably defined clinical complications) introduce an important source of possible information or measurement bias and because of their heterogeneity, confuse the clinical utility due to lack of a clear target population. Moreover, heterogeneity of treatment definitions and outcomes assessed as well as lack of rigorous confounding adjustment make it difficult to draw valid inferences from published summaries of these data\textsuperscript{39-41}. Two systematic reviews\textsuperscript{40, 41} qualitatively summarized this literature and reported inadequate evidence to suggest difference in morbidity or mortality between early (usually less than 24 hours) versus late operative treatment of femoral shaft fractures, overall or within subgroups of patients with associated head injuries or thoracic injuries. These inconsistent findings have left clinicians debating what causal effect definitive fracture care has on adverse outcomes.
Problems with treatment definition are numerous. Treatment time has typically been dichotomized so as to compare fixation before to fixation after twenty-four or forty-eight hours (most commonly used cut-points). While categorization of a continuous measure such as time helps to draw clinically relevant distinctions, a simple dichotomy needlessly limits the surgeon’s decision-making possibilities when more choices to represent this continuum would better match reality. Second, the term treatment has come to mean many different things in the literature. Early studies assessing the effect of treatment time tended to compare early medullary nail fixation of femur fractures to late fixation or non-operative fixation (i.e. skeletal traction) and tended to conclude that early fixation was beneficial1, 10, 16. In the modern era of orthopaedic trauma care, non-operative management of a femoral shaft fracture is not a therapeutic option and fractures are either definitively stabilized early or late by internal fixation (plate osteosynthesis or medullary nail fixation) or external fixation. This is done to allow early mobilization in order to prevent the complications of long-period recumbence including venous thromboembolic disease, pneumonia and decubitus skin ulcerations. Therefore, the conclusions of many studies may have been mixing the effects of operative management with any independent effects from treatment time.

Outcome assessment and reporting has been equally problematic for discerning possible treatment effects of choice of timing. Measures of outcome have ranged from the objective and rare (i.e. mortality) to the subjective and unreliably reported (Acute Respiratory Distress Syndrome, fat embolus syndrome, or pneumonia). Small sample sizes have led to inadequate power to detect differences in mortality forcing the majority of published studies to follow reported complications that often lack systematic or unbiased assessment. Several investigators have turned to reporting differences in length of stay (LOS) or some component of LOS (i.e. time spent on a ventilator or in the intensive care unit) as a surrogate for morbidity caused by the timing of definitive fracture care. Unfortunately, the causal interpretation of treatment time effect on LOS is complicated by whether tabulation starts from the time the patient enters the hospital and a delay in treatment is chosen, or from the time of definitive treatment. Arguments can be made in support of either approach, though most studies have taken the former approach and looked at total hospital and intensive care unit stay. The use of LOS is also subject to vagaries of clinical and institutional practice such as insurance-related contracts for transfer of patients to sub-acute levels of care and selection bias due to censoring by mortality. Without an objective or reliable method of outcome measurement, the effect of treatment time has been difficult to fully assess.

Confounding bias has been one of the greatest threats to valid inference from the data that have been reported on this subject. In any study of the causal effects of treatment, the potential for confounding by indication for exists42, 43. In the case of multi-system trauma, patients who are treated early typically have a better prognosis given less severe injuries than those who are treated later. This has the potential to bias results towards favoring early fixation. This phenomenon of confounding by indication has either been ignored or addressed with only crude methods of statistical adjustment in most studies on this subject to date.

The literature has also struggled with how to identify those patients who may be most susceptible to the effects, positive or negative, of early treatment. This is fundamentally a question of interaction where-by an investigator strives to understand whether the treatment of interest has potentially different effects among subgroups of patients defined by the presence or
absence of some characteristic. Tests for homogeneity of effects across strata of such extraneous characteristics have been classically described conditional on treatment or exposures received using stratification\textsuperscript{44, 45} as well as conventional regression\textsuperscript{46}. Unfortunately, with regards to this recurrent investigation of patients with and without serious chest injury or head injury, no formal assessments of interaction have been made at all. Instead, improper comparisons of incidence of an outcome of interest have been compared. For example, in the stratified analysis performed by Pape\textsuperscript{25} discussed earlier, these authors reported a significant increase in the incidence of ARDS among chest injured patients who were treated with medullary nailing early (<24 hours from admission) versus those who were not chest injured (33% versus 7.7%) and a similar but not statistically significant difference in mortality (21% versus 4%). While these authors infer interaction by presence of chest injury, their study fails to contrast the effect of treatment time between subgroups defined by this associated injury. Ultimately, even well established methods for assessing interaction are conditional on treatment received by patients. The ideal test for interaction would be a comparison of the causal effects of treatment among all patient with or without a given treatment within strata of factor of interest, a parameter that has until recently been undefined\textsuperscript{47} and will be discussed in further detail in Chapter 3.

The multidimensional nature and urgency of treatment assignment have made it logistically difficult and ethically dubious to conduct randomized clinical trials to evaluate the effect of treatment timing on patient outcomes. Observational studies are therefore the primary source of data for the investigation of this research question. Yet the appropriate means by which to analyze such data for causal effects of early treatment have not been established. The majority of published studies have either reported crude (unadjusted) estimates or stratified their analyses by one or a handful of potentially confounding factors such as ISS or age. A few studies have used conventional regression techniques such as linear and logistic regression that “adjust” for one covariate at a time and ignore the joint distribution of confounding. Moreover, these “conventional” methods offer conditional associations that are difficult to interpret in the face of numerous covariates and interaction terms and will be reviewed in the following chapter. Hence, the observational nature of reports to date and their limited consideration of confounding severely limit any causal inferences that are desired by investigators and clinicians who look to them for answers.

**Study question**

In an attempt to address methodological limitation discussed, I conducted a study with specifically defined parameters and analytical methods for estimation of causal effects with reference to the following two specific aims:

1. To estimate the association of early definitive stabilization of the femoral shaft fracture on in-hospital mortality in multi-system trauma patients via semi-parametric, “causal inference” methods.

2. To assess the causal effect of early definitive stabilization of the femoral shaft fracture on mortality among subgroups defined by the presence or absence of associated serious head, chest or other skeletal injury in order to identify those patients that are most sensitive to the effect of treatment time.
Chapter 2 : Conventional and Semi-parametric Approaches to Confounding and Covariate Adjustment in Observational Studies

Properly conducted randomized controlled trials (RCT) offer the highest level of evidence from clinical research investigating therapeutic interventions. Yet there are many scenarios where RCTs are inappropriate or impossible such as studies of rare conditions or those with prolonged induction or latent periods; moreover, the generalizability of RCTs is often limited by strict inclusion and exclusion criteria. Non-randomized or observational studies can provide an important complementary source of information when RCTs cannot or should not be undertaken, provided that the data are analyzed and interpreted in the context of the many forms of bias to which they are prone. Clinicians and researchers trying to understand the literature pertaining to questions such as the optimal timing for femoral shaft fixation are inundated with claims of association of risk factors and assertions of benefit of certain treatments for a given condition. Understanding study design strategies and statistical methods of bias reduction are essential to interpreting data and determining what constitutes useful evidence.

In this chapter I will focus on causation and confounding bias in the context of observational orthopaedic outcome studies and discuss strengths and weaknesses of various design and analysis strategies that have been undertaken. I will start by discussing basic principles including relationship between a study sample and the target population. The notion of probability distributions will be presented as a means of understanding the two primary forms of statistical analysis, estimation and hypothesis testing. The counterfactual approach to causation will be introduced as a framework to understanding threats to valid inferences and examples will be discussed from the orthopaedic literature as well my own research to illustrate concepts.

Basics Principles of Statistical Inference

The analysis of any clinical study is based on the principle of taking a sample of subjects for the purposes of drawing inferences about a larger population of similar individuals (Figure 2.1). Taking a carefully selected sample that is randomly selected or representative of the population of interest is a powerful way to make inferences about the target population. However, going from population to a sample leads to some degree of uncertainty or margin of error because we are estimating without knowledge of entire population. We rely on mathematically defined probability distributions in order to quantify this uncertainty such as the Normal distribution for continuous data and Binomial distribution for categorical data. These distributions are based on parameters such as a mean and standard deviation. If the assumption is made that the observed data are a sample from a population with a distribution that has a known theoretical form, then it is reasonable to use parameters of that distribution (those observed) to calculate probabilities of different values occurring. This parametric approach to statistics is wide-ranging and ubiquitous in medical research. However, if these distributional assumptions are not realistic, then we may end up with results that are not valid. When data deviate from a so-called “Normal” pattern, non-parametric (or semi-parametric if there are some aspects of the distribution that can be assumed known) or distribution-free methods are required.

6
**Estimation and hypothesis testing.** One of the primary objectives of any clinical study is to provide some numerical value expressing risk of outcome or of relative effect associated with a specific treatment or prognostic factor. Estimation covers a broad range of statistical procedures that yield the magnitude of risk or effect as well as the precision or confidence interval of that estimate. Hypothesis testing concerns itself with understanding the likelihood of having observed a difference or association from data if no such effect exists in the population. The ubiquitous use of this likelihood, known as the P-values, leads to treating the analysis as a process of decision-making within which it is customary to consider a statistically significant effect as real and non-significant result to indicate no effect. Notice that this value gives no information about the magnitude of interest or the uncertainty of the result. For these reasons the approach based on estimation is often considered superior.

The threshold level at which one may consider a P-value to be statistically significant also depends on how many times one sample group is compared to another. The more times that one looks for a difference between two groups, the more likely one is to find a difference there that has occurred purely by chance (i.e. the Type I error rate increases). This is an important consideration when two groups defined by treatment received are compared with respect to multiple outcomes of interest, or multiple subgroup analyses are performed. This is ideally done as a way of using the data to generate new hypotheses for future study, rather than trying to use the same data to try to definitively answer multiple questions. When multiple comparisons are made and multiple hypothesis tests undertaken with the same data, the cut-off P-value for statistical significance should be lowered accordingly using a Bonferroni or similar adjustment method to guard against Type I error.

**Bias, Confounding, and Counterfactuals**

Gathering useful information from non-randomized studies requires a clear understanding of the role of bias in the data and how it aught to be handled. While there are multiple described forms of bias that threaten the validity of a clinical research study, most fall into one of three categories: selection bias, information bias, and confounding. Selection bias is defined as a distortion of estimates that results from the way in which subjects are selected into the study sample and includes factors such as differential inclusion or diagnosis of persons in the study and loss to follow-up. For example, if likelihood of censoring is influenced by which treatment or implant a patient receives, analyzing only those that followed up will bias any relative estimate of effect. Information bias arises from systematic or non-systematic difference between groups in the way that data are obtained. This is a bias that may result from a non-blinded surgeon assessing outcomes of a particular treatment and, like selection bias, is also a threat to validity no matter what study design is used. Confounding bias represents a mixing of effects between the treatments of interest and associated extraneous factors that also impact outcome, potentially obscuring or distorting the relationship of interest. Confounding is of particular concern in observational studies, and surgical research specifically when causal relationships are investigated.

Despite the importance of understanding and proving causal relationships in scientific research, a formal definition has been debated for centuries. Greenland has reviewed this chronology from the early definitions of the eighteenth century Scottish philosopher, David Hume through a statistical derivation based on potential outcomes by the early twentieth century statistician, Jerzy Neyman. These authors have established a causal exposure (A) as necessary...
for the occurrence of a measured outcome (Y) under observed background circumstances such that had the cause (A) been altered, the effect (Y) would have changed. A causal effect, then, is a contrast between the outcomes of a single unit under different treatment possibilities (y_{ik} versus y_{i0}). However, a single unit can only receive one treatment so such a contrast (y_{ik} - y_{i0}) can never actually be made. Still, this theoretical framework using an unobservable comparator group provides the basis for the counterfactual approach to causality.

In empirical research, investigators can only sample from observed groups (say A and B) that receive different treatments (A = 0 or A = 1) and compare the observed mean outcomes \( \mu_A = E(Y=1|A=1) \) in group A, and \( \mu_B = E(Y=1|A=0) \) in group B (see Table 2.1). A causal interpretation from the observed comparison requires that \( \mu_A - \mu_B \) equal \( \mu_A - \mu_B \) and therefore, \( \mu_A \neq \mu_B \). This implies \textit{exchangeability} in the response distributions under homogeneous treatment assignment. This assumption is satisfied generally when treatment assignment is independent of outcome, the successfully randomized controlled trial being the ideal example of such a scenario. In RCTs, the randomization process will, on average, evenly balance both known and unknown confounders, and this guarantees the validity of the statistical test used. The randomization process makes it possible to ascribe a probability distribution to the difference in outcome that is not influenced by differences in any prognostic factor other than the intervention under investigation\textsuperscript{58}. The chi-squared test for 2-by-2 tables and the Student’s t-test for comparing two means can be justified on the basis of randomization alone without making further assumptions concerning the distribution of baseline variables. In the absence of randomization, additional design and analysis methods are needed to account for those sources of bias arising from a lack of comparability or \textit{exchangeability} between groups. Unfortunately, most these methods only account for known sources of bias and come with there own set of assumptions that can only partially be tested.

As mentioned earlier, confounding is a particular problem in non-randomized studies. Confounding can arises when patients selected for one treatment group are fundamentally different from the other group in their pre-treatment likelihood for the outcome of interest. In the notation introduced above, confounding is present when \( \mu_{A0} \neq \mu_{B0} \). A patient characteristic or associated factor may be considered a confounder when: 1) it can causally affect the outcome parameter within treatment groups; 2) it is distributed differently among compared populations; and 3) is absent from the causal pathway of interest. It is important to note that confounding depends not only on single covariate differences between compared groups, but in the joint distribution of confounding factors. In surgery, treatment decisions are commonly made on grounds of certain overt and subtle factors related to prognosis or severity of disease. As discussed in Chapter 1, if patients with less severe systemic injuries are able to undergo fixation of their long-bone fractures at an earlier time than those with severe multi-system trauma, there may be factors other then timing of treatment that influence the lower rates of morbid complications among those treated early. This type of bias in studies of therapies has been termed \textit{confounding by indication} or \textit{confounding by severity}\textsuperscript{59} and is a major threat to validity of conclusions from non-randomized studies.

\textit{Controlling Confounding}

Control of confounding can be undertaken either in the design phase or the analysis phase of a study, or in both. Design-based methods (randomization, restriction, matching) attempt to
make $\mu_{A0} = \mu_{B0}$ in the study sample and analysis-based methods (stratification, regression) typically assert this equality but only within strata of a confounder or regressor. These methods as well as two more advanced multivariable methods that have been increasingly used in analysis of therapeutic studies, propensity score analysis and instrumental variable analysis, are discussed below. Restriction, matching, stratification, regression, and even propensity score-based methods work through the notion of fixing the level of one, or more, factors constant in order to study the variability in outcome specifically associated with a change in the treatment or risk factor of interest. The interpretation of such analyses is therefore appropriately described as conditional on holding other known variables constant (again, with the exception of instrumental variable analysis). Achieving sufficient control of confounding for estimation of stratum-specific causal effects is contingent on exchangeability of $\mu_{A0}$ and $\mu_{B0}$ within strata of the confounder or confounders. Randomization assures sufficiency of confounding control through its probabilistic balancing of known and unknown confounders as described above. In the absence of randomization, causal inference becomes contingent on the identification of a set of baseline factors whose control will be sufficient. This is based on a priori subject matter knowledge and can be systematically illustrated in causal diagrams that depict the direction of causal relationships and allow for identification of minimally sufficient subsets of variables for control. Figure 2.2 is a potential causal diagram depicting important variable relationships for estimation of the causal effects of treatment time on mortality in multi-system trauma patients with femoral shaft fracture. Restriction is simply an attempt to enforce homogeneity within the study sample by only including those subjects with the same value of one or more potential confounders, thereby removing any potential associations with outcome. This is probably the most effective approach in removing confounding by any one variable, however it estimates a different parameter than generally desired (i.e., only within a target population defined by the restriction) but the curse of dimensionality for high-dimensional data (lots of covariates) will render such an approach unfeasible. Far more commonly used in the orthopaedic literature are the use of matching, stratification, and multivariable regression.

Matching is a conceptually straightforward strategy whereby confounders are identified and subjects in the treatment groups are matched based on these factors so that they are in the end, “the same” with regards to these factors. Matching can either be done on a one-to-one basis (matched pairs), or based on frequencies (i.e. equal percentage of subjects in each group with confounder) and subjects can be matched with respect to a single confounder or multiple confounders. Matching can be used in both prospective and retrospective non-randomized study design (including case-control studies). For example, Ciminiello examined the impact of small incisions (<5cm) on a variety of outcomes including blood-loss, operative time, and post-operative complications in patients undergoing primary total hip arthroplasty (THA). In order to assure that the group of patients undergoing THA with small incision and the comparator group undergoing THA with a standard size incision were as homogenous as possible, the authors used a matched-pair cohort design. They matched 60 patients in each group on a variety of potentially confounding factors including age, sex, body mass index, American Society of Anesthesiologists Score, diagnosis (osteoarthritis), prosthesis, type of fixation, anesthesia, surgical approach and positioning and were unable to identify any significant differences in outcome between the two techniques. Matching is a useful method for improving study efficiency, especially with such known confounders measured on a nominal scale.
While matching is an effective way of balancing multiple confounders, it has several important limitations to consider. One is that it may be difficult or impossible to find exact matches between the two groups of patients, and this difficulty increases rapidly as the number of factors that one wishes to match on increases. While a better understanding of the relationship between baseline variables, including treatment, and outcome may come from an analysis of the entire cohort, matching may sacrifice significant numbers (and therefore, power) to study the relationships of interest (in essence, it capitalizes on no smoothness in the conditional distribution of the outcome given the covariates, allowing no “borrowing” across different strata via a model). One solution to this is matching within a reasonable range (for example, increasing this “borrowing” by allowing age \( \pm 5 \) years) within which differences of prognostic importance are not thought to exist. Another problem is that matching generally precludes the evaluation of the underlying relations between matching variable and exposure in a prospective cohort study and matching factor and outcomes in a case-control study. This is because of the sampling schemes (based on exposure for the prospective cohort study, and outcome for the case-control) and the way that balance is forced with respect to the matching factor in each of these designs. Finally, if matching is undertaken on variables that are not true confounders, a loss of statistical power can result; moreover, in a case-control study such “overmatching” can create a new bias. Therefore, matching should be used cautiously and only on factors that are strong risk factors for the outcome of interest and thought to be differentially distributed between treated and untreated subjects (i.e. only on true confounders).

*Stratification* (one could consider as a less severe form of matching) provides another means by which to control confounding. Potentially confounding variables are identified and the cohort grouped by levels of this factor. The analysis is then performed on each subgroup within which the factor remains constant, thereby removing the confounding potential of that factor. If the estimates among stratified groups are relatively homogenous, they can be averaged into an estimate unconfounded by the stratification variable, usually by the Mantel-Haenzel method. Bosse undertook a study to assess the impact of reamed intra-medullary nailing versus plate fixation of patients with femoral shaft fractures on several adverse outcomes including adult respiratory distress syndrome and multiple organ failure. Because the presence of thoracic injury could cause differences in surgical approach and impact the outcome of interest, this was considered a potential confounder and stratified analysis was undertaken. Table 2.2 shows the crude or unadjusted analysis of the cohort as well as the subgroup analysis stratifying on presence or absence of thoracic injury. The stratified analysis shows that confounding by thoracic injury did not significantly alter the relative risk of complication seen in the crude analysis. Because a very fine level of stratification would lead essentially to matching (treated to untreated), the use of the term relative to matching is somewhat arbitrary.

Stratification is a useful strategy when there are only one or two risk factors or confounders that allow easy grouping. It also allows for easy visualization of interaction or a difference in treatment effect across subgroups defined by a separate factor. For example, if there were a difference in the effect of surgical treatment on complications depending on whether or not patients had thoracic injury in the study by Bosse, this difference itself could be tested and reported as a separate finding of interest. The problem with stratified analysis (much like occurs in matching on a large number of potential confounders) is that it quickly becomes unmanageable and difficult to interpret when there are multiple confounders with multiple levels each. In this scenario, it is better to turn to multivariable regression.
Multivariable Regression for the adjustment of confounding factors is one of the most common techniques used in therapeutic and prognostic observational studies. Regression analysis is based on modeling the mathematical relationships between two or more variables that give an approximate description of the observed data. Regression models should not be thought of as explaining underlying mechanisms (i.e. statistical models are not reality), but rather as simplifications that are compatible with the data and provide us with some inference as to associations seen in the data. These models are usually additive in that an observed dependent variable (i.e. outcome of interest) can be explained by a model in which the effects of different influences or independent variables (i.e. treatment of interest as well as other predictors of outcome or confounding factors) are added. Though these models are not based on theory, they should in principle be fit in a manner that “learns” from data. However, most analyses are based on some variation of an arbitrarily chosen the generalized linear model:

\[ h(E[Y|X]) = \alpha + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_pX_p \]

where a function, \( h \), of the conditional expectation (or mean value) of \( Y \) given \( X \) (or \( E[Y|X] \)) is an additive combination of an intercept (\( \alpha \)) and (\( p \)) explanatory independent variables multiplied by their respective coefficients (\( \beta_1 \) through \( \beta_p \)); \( h \) is referred to as a link function and includes the identity link (linear regression), log and logit (logistic regression). Each coefficient represents an estimate of effect or risk depending on \( h \) (Examples: mean difference for linear regression and log odds ratio for logistic regression). Multivariable analysis allows one to estimate the association between dependent and independent variables, controlling for influence of other independent variables.

The exact model that is used typically depends on the type of outcome data one has. For example, Saleh\(^63\) performed a case-control study to evaluate the predictors of surgical site infections (a binary outcome) complicating total knee and total hip replacement. These authors used multivariable logistic regression to control for several demographic, peri-operative, and postoperative factors and found post-operative hematoma formation and persistent wound drainage to be the only significant associated risk factors. In a study looking for prognostic factors associated with re-operation after operative treatment of tibial shaft fracture, Bhandari\(^64\) used a proportional hazards analysis in order to assess multiple risk factors. After adjusting for over twenty possible variables, three predicted a relative increase in re-operation (open fracture, transverse pattern, and lack of cortical continuity). This form of analysis gives an estimate of effect that is analogous to the odds ratio from logistic regression called the hazard ratio, and can similarly be interpreted as a relative measure of risk of event associated with a unit change in a given predictor, holding other factors constant.

While the results of such multivariable analysis are commonly presented and accepted as absolute, the details of how the models were selected are rarely scrutinized. Readers may be led to assume the results are accurate when they may have been derived using inappropriate models. Control for confounding solely by regression obscures the ability to visualize the adequacy of overlap or exchangeability between treatment and control groups and relies on dubious model-based extrapolations. Regression models assume that there is no effect modification, or difference of effect between different levels of a confounder as discussed earlier in the stratification example. Unless an interaction term (usually a product of two predictor variables) is inserted to represent interaction, the model may remain naïve to such a relationship and falsely
conclude that the effect of treatment A on outcome Y is constant across all levels of factor W (i.e. gender). Assumptions are made when a model is fit (such as that of multivariable Normality for linear regression or that the relative contribution of each factor is constant over time for a proportional hazards model) and it is important that these assumptions are verified and the overall fit of the model to the data be assessed. Model fit is determined in terms of the amount of variability in the data explained by the model and how well the model predicts individual outcomes for a given observation. There are many “diagnostic” procedures for assessing the most valid and best fitting regression model\textsuperscript{65, 66}. Unfortunately, the phenomenon of “wrong model bias” is underappreciated in the orthopaedic literature as is it in epidemiological analysis in general. Given the availability of semi-parametric methods, machine learning algorithms and interesting association parameters that do not depend on the arbitrary specification of a specific regression model, there is enormous improvement possible in the standard of practice by both choosing interesting relevant parameters and leveraging the information contained in the data, while not having the results tainted by arbitrary (and often consequential) assumptions on a model with no theoretical or empirical basis.

The multidimensional nature and urgency of treatment assignment in the acute trauma setting have made it logistically difficult and ethically dubious to conduct randomized clinical trials to evaluate the effect of treatment timing of femoral fracture on patient outcomes. Observational studies are therefore the primary source of data for the investigation of this research question. Yet the appropriate means by which to analyze such data for causal effects of early treatment have not been established. Conventional regression techniques such as linear and logistic regression that “adjust” for one covariate at a time ignore the joint distribution of confounding. Moreover, these “traditional” methods offer conditional associations that are difficult to interpret in the face of numerous covariates and interaction terms. Propensity score analysis has features addressing some of these problems.

Propensity score analysis\textsuperscript{67} is another approach to controlling for confounding through the generation of a score that “summarizes” the confounding by multiple variables. The propensity score is defined as a balancing score that is a function \( b(X) \) of the covariates \( X \) such that the conditional distribution of \( X \) given \( b(X) \) is the same for treated \( (A=1) \) and control \( (A=0) \) units. This form of analysis requires a two-stage approach in which first, rather than modeling the outcome as a function of multiple risk factors, the probability of being treated is modeled taking into account any possible confounding variables. This probability, possibly generated by a logistic regression model, is the propensity score and ranges from 0 to 1. Once the propensity score is generated for each subject, it can be used to match them (usually within some narrow range), perform stratified analysis on levels (such as deciles) of the propensity score, or it can be inserted into multivariable regression along with the treatment variable in estimating the outcome. Propensity score analysis tends to produce unbiased estimates of the treatment effects when treatment assignment is strongly ignorable (treatment assignment is considered strongly ignorable if treatment assignment, \( A \), and the response, \( Y \), are known to be conditionally independent given the covariates, \( X \)). More simply stated, this is the assumption of no unmeasured confounding.

While orthopedic investigators have been slower to apply propensity score analysis than researchers in the medical specialties such as cardiology and cardiothoracic surgery fields in particular\textsuperscript{68}, there are some examples. McHenry and colleagues\textsuperscript{69} used propensity score analysis
to control for confounding by indication for timing of treatment of surgically treated thoracic and lumbar spine fractures in order to assess risk factors for respiratory failure following operative stabilization of these injuries (i.e. this was a prognostic study but used propensity score methods to adjust for the strong likelihood of differential treatment-time assignment due to injury severity). Subjects were matched based on their propensity score for treatment before 48 hours from injury. Logistic regression was then performed on the matched set, identifying age, Injury Severity Score, Glasgow Coma Score, presence of blunt chest injury, and timing to surgery more than 2 days as independent risk factors for respiratory failure. By matching on the propensity score, these investigators were able to limit bias due to an important surgeon controlled risk factor in assessing the relative importance of multiple prognostic factors.

By estimating the treatment mechanism, propensity score analysis offers several insights into the data and theoretical advantages over conventional techniques of multivariable adjustment. First, propensity scores allow one to see the degree to which likelihood of treatment differs between two groups and allows one to assess how comparable treatment groups actually are (i.e. the two groups should have fairly similar distributions of propensity scores to make the comparison tenable). Preliminary analysis from my study, further described in Chapter 3, showed qualitatively similar distributions of the estimated propensity score among multi-system trauma patients whose femoral shaft fractures were treated before and after 12 hours from admission (Figure 2.3). Second, by matching or stratifying subjects on the basis of their likelihood of treatment, one can intuitively see how selection bias is countered because comparisons are made only among those equally likely to have received treatment, as in a randomized controlled trial. From preliminary analysis just mentioned, stratification by decile of the estimated propensity score shows remarkable similarities within strata for at least three variables thought to be strong confounders of the treatment time-mortality relationship (Table 2.3).

Other advantages of this strategy over standard regression include confounding control that is more robust to modeling assumptions because one can be less parsimonious about numbers and combinations of terms to optimize model fit in estimating the propensity score and propensity score analysis has shown superior performance in the setting of low numbers of events per confounder adjusted for (less than eight events per confounder). Unfortunately, like all but instrumental variable methods, propensity score analysis is no more immune to threats to validity caused by unknown or unmeasured confounders. Also, two recent systematic reviews have not shown significant differences in estimates from studies performing side-by-side conventional multivariable and propensity score analyses. While the use of propensity score analysis is growing quickly among many fields of research in medicine, guidelines for proper use of this methods are lacking. In addition, because the outcome is not used in the selection of the model for the propensity score, the approach is inherently inefficient (e.g., consider a variable which is a strong predictor of treatment but unrelated to the outcome, which is sure to be selected in the propensity score model and yet damage the efficiency of the resulting estimator while not contributing to reduction in the bias). In addition, because only the untreated can be matched with a treated subject, under assumptions the propensity score estimates a specific parameter, which does not apply to the entire target population but instead the association only among the treated. Distance algorithms are equivalent alternatives to matching on the propensity score.
The *instrumental variable* approach to bias and confounding in medical research has been used frequently by econometricians for decades but only recently been implemented in health research\textsuperscript{78-80}. Health economists now commonly use this method to examine questions about quality and distribution of care when using administrative data where many confounders are potentially missing. The theoretical advantage of instrumental variable methodology in the analysis of non-randomized therapeutic studies is that it offers the possibility of controlling of both known *and* unknown confounding and is therefore appealing when the threat of unobserved or unobservable confounders looms large. The idea is as follows: if a variable (the instrumental variable) can be identified that has the ability to cause variation in the treatment of interest and has no impact whatsoever on outcome (other than that which passes directly through its association with treatment), then one can estimate how much the variation in the treatment is induced by the instrument and only that induced variation affects the outcome. Figure 2.4 provides a schematic diagram of this prerequisite relationship for the identification of a useful and valid instrument. Instrumental variables can be thought of as achieving pseudo-randomization. The RCT is a special case where the random number assignment (say a fair coin toss) is the instrumental variable inducing variation in the outcome variable.

Like propensity score analysis, examples of the use of instrumental variable analysis are rare in the orthopaedic literature. McGuire and colleagues\textsuperscript{81} used a large Medicare data set to examine the controversial topic of impact of timing of fixation of hip fractures on mortality. Acknowledging the fact that there are likely prognostic factors beyond those measured in such a large administrative data set, these investigators chose day of the week grouping (Saturday through Monday vs. Tuesday through Friday) as an instrumental variable by which to pseudo-randomize the cohort. The authors site evidence showing that day of the week is a strong predictor of delay to operative hip fracture treatment and assume that day of the week that one breaks one’s hip should have no independent influence on mortality or association with other confounders such as presence of co-morbidities. The instrumental variable analysis showed an increased risk of mortality (risk difference 15%, p=0.047) among those undergoing surgery more than two days from admission. It is likely that this methodology will grow in popularity among researchers trying to draw unconfounded estimates of effects of similar health care decision from large datasets for clinical practice and policy-making reasons.

While the thought that one can get around unobserved (and therefore uncontrolled for) confounding in non-randomized studies is appealing, there are certain important limitations for using these methods to establish causality. First, indentifying an instrumental variable that meets the assumptions of no association with outcome, independent of treatment, is difficult. Because this assumption is not directly testable, there must be general consensus that the instrumental variable is tenable. In comparing instrumental variables with standard multivariable adjustment or propensity score techniques, one is trading the assumption just mentioned for the assumption of no unmeasured confounding which is also not directly testable. The other important consideration when interpreting the result of an instrumental variable analysis, assuming the instrumental variable has been chosen correctly, is that the effect measured only applies to those whose treatment was affected by the instrumental variable. In the study by McGuire\textsuperscript{81}, 15% increase in risk of mortality associated with delay of surgery applies only to the patient whose treatment timing was influenced by the day of the week they were admitted. This so called *marginal patient*\textsuperscript{82} within a cohort study, is important to distinguish from the entire study sample to whom any average treatment effect can be inferred in a randomized controlled trial. In general,
there will also be other required restrictions (such as no non-compliers among the untreated) and these methods, because they are estimating association parameters in a very large model, require enormous sample sizes to be practical statistical tools. Thus, their popularity in econometrics where enormous populations are often measured for the outcomes and variables of interest.

Comments

Non-randomized studies continue to provide an important method for clinical investigation in orthopedic surgery in settings were randomized controlled trials are not feasible and when increased generalizability of findings is desired. Confounding bias represents a major obstacle to drawing valid inferences and has been described here from a counterfactual approach to causal inference. Design and analysis tactics for control of confounding have been elaborated on as to their respective strengths and weaknesses. Restriction, matching and stratification are useful but inefficient means when dealing with multiple confounding factors. Conventional multivariable adjustment offers the power to adjust for multiple confounders at the same time, but cannot give causation unless factors like appropriate temporal ordering of predictor and outcome are assured, and there are no unaccounted for confounders in the analysis. While propensity score analysis offers a more plausible accounting for the multivariable nature of confounding and balancing of confounding by indication, causal interpretation from such is still limited by the similar assumptions. Instrumental variable analysis depends on universal acceptance of the instrumental variable. The ensuing chapters will describe an approach novel to the orthopaedic surgery literature, using marginal structural models to explore causal effects of treatment time on mortality in multi-system trauma patients with femoral shaft fracture.
CHAPTER 2 FIGURE LEGEND:

Figure 2.1: Target population, study sample, and statistical inference. (Reprinted with permission, Journal of Bone and Joint Surgery-American Edition)84

Figure 2.2: Causal diagram depicting the relationships between treatment time of multi-system trauma patients with a femoral shaft fracture and mortality as well as known and unknown potentially confounding covariates.

Figure 2.3: Distribution of the propensity score among the early (<12 hours) and late (≥12 hours) treated groups of multi-system trauma patients with femoral shaft fracture. Propensity score estimated using a logistic model including Injury Severity Score, Glasgow Coma Scale, age, time of arrival, number of serious extremity/pelvic or head/neck injuries, number of femoral fractures, presence of cardiac or cerebro-vascular co-morbid condition, hospital teaching status, American College of Surgeons Level-1 designation.

Figure 2.4: Assumptions of an appropriate instrumental variable (IV): 1) IV must be associated with treatment; 2) IV must have no association with outcome, other than through its influence on treatment. * Confounders here represent both those observed and unobserved. (Reprinted with permission, Journal of Bone and Joint Surgery-American Edition)84
Figure 2.1
Variable Definitions

- A = Treatment Time
- Y = Mortality
- W = Confounding Covariates
- U = Unknown or Unmeasured Confounders

Figure 2.2
Figure 2.3
Figure 2.4
Table 2.1: Contingency table of observed outcome (Y) among groups of individuals defined by treatment (A). \( \mu_{A1} \) is the mean outcome or incidence in those receiving treatment (A=1), \( \mu_{B0} \) is the mean outcome in those not receiving treatment (A=0).

<table>
<thead>
<tr>
<th></th>
<th>Y=1</th>
<th>Y=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>a/(a+b)= ( \mu_{A1} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A=0</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>c/(c+d)= ( \mu_{B0} )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Table 2.2: Stratified Analysis of Nail versus Plate Fixation and Development of Adult Respiratory Distress Syndrome or Multiple Organ Failure from Bosse et al\textsuperscript{17}

<table>
<thead>
<tr>
<th>Chest Injury</th>
<th>No Chest Injury</th>
<th>Total Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Plate</td>
</tr>
<tr>
<td>Developed Adult Respiratory Distress Syndrome or Multiple Organ Failure</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Number At Risk</td>
<td>117</td>
<td>104</td>
</tr>
<tr>
<td>Risk</td>
<td>0.043</td>
<td>0.019</td>
</tr>
<tr>
<td>Risk Difference (95% Confidence Interval)</td>
<td>0.024 (-0.022, 0.069)\textsuperscript{*}</td>
<td>0.025 (-0.012, 0.062)\textsuperscript{*}</td>
</tr>
<tr>
<td>Summary Risk Difference (95% Confidence Interval)</td>
<td>0.024 (-0.005, 0.053)\textsuperscript{\textdagger}</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{*}Risk Differences between strata are not significantly different, that is, no interaction (Test for Heterogeneity P-value=0.96).

\textsuperscript{\textdagger}Given the absence of interaction, a pooled summary validly estimates the Risk Difference adjusting for Chest Injury.

\textsuperscript{+}P-value testing the null hypothesis of no association between treatment method and outcome, stratifying by chest injury and assuming no interaction.
Table 2.3: Balancing of three covariates highly associated with both treatment time and mortality among the study sample of multi-system trauma patients undergoing femoral shaft fixation. GCS – Glasgow Coma Scale; Age (years); ISS – Injury Severity Score. Columns represent deciles of the propensity score. Propensity score estimated using a logistic model including Injury Severity Score, Glasgow Coma Scale, age, time of arrival, number of serious extremity/pelvic or head/neck injuries, number of femoral fractures, presence of cardiac or cerebro-vascular co-morbid condition, hospital teaching status, American College of Surgeons Level-1 designation. E - Early treatment (<12 hours from admission); L – Late treatment (≥12 hours from admission). Cells contain the mean value for respective variable at that decile of the propensity score and treatment assignment (Early or Late). Treatment groups compared for each variable at each decile using the Wilcoxon rank-sum test. Highlighted cells represent the only comparison with a test statistic p<0.05.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCS</td>
<td>28</td>
<td>95</td>
<td>85</td>
<td>142</td>
<td>96</td>
<td>127</td>
<td>129</td>
<td>123</td>
<td>140</td>
<td>128</td>
</tr>
<tr>
<td>Age</td>
<td>38.6</td>
<td>39.5</td>
<td>35.3</td>
<td>35.4</td>
<td>35.9</td>
<td>31.5</td>
<td>33.2</td>
<td>34</td>
<td>32.2</td>
<td>31.9</td>
</tr>
<tr>
<td>ISS</td>
<td>36</td>
<td>36</td>
<td>29</td>
<td>29</td>
<td>27</td>
<td>27</td>
<td>26</td>
<td>26</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Highlighted cells represent the only comparison with a test statistic p<0.05. |     |     |     |     |     |     |     |     |     |     |
Chapter 3: Estimation of marginal structural models to estimate causal effects of timing of treatment of femoral shaft fracture

The prior chapters have introduced the study question in the context of extant literature and provided an overview of confounding bias and methods currently used to minimize it in observational studies of orthopaedic interventions. In this chapter I explain an application of marginal structural models to study the effect of treatment timing on mortality in multi-system trauma patients.

Study Cohort and Data Source

I derived the study cohort from the National Trauma Data Bank (NTDB) version 5.0 which draws from 567 trauma centers from around the United States\textsuperscript{45} and contains nearly one million incident trauma cases that occurred over a five year period between January 1\textsuperscript{st}, 2000 and December 31\textsuperscript{st}, 2004. The NTDB contains pre-hospitalization and hospitalization data compiled from medical records during admission—including injury characteristics, co-morbidities, inpatient treatments, and outcomes—that are submitted to the American College of Surgeons for quality control and maintenance.

Patients were included in the study sample if they: 1) had closed or open fracture(s) of the femoral shaft identified by an International Classification of Diseases, 9\textsuperscript{th} Clinical Modification (ICD-9CM) diagnostic code of 821.0 (fracture femur shaft/not otherwise specified—closed), 821.01 (fracture femur shaft—closed), 821.1 (fracture femur shaft/not otherwise specified—open), or 821.11 (fracture femur shaft—open); and 2) had an International Classification of Diseases - derived Injury Severity Score (ISS)\textsuperscript{18,86} of greater than or equal to 15; and 3) were 16 years of age or older; and 4) underwent a “definitive” treatment procedure involving internal fixation of the femur identified by an ICD-9CM procedure code of 78.55 (internal fixation—femur), 79.15 (closed reduction—internal fixation femur), or 79.35 (open reduction—internal fixation femur). Patients were excluded from the study sample for any of the following reasons: 1) the patient was received in transfer or not admitted on the day of injury; 2) the record lacked information on time from admission to definitive fracture fixation, mortality status, or length of hospitalization; 3) there was an associated burn; or 4) the patient did not have the fracture definitively fixed within two weeks of admission.

Because of the potential for treatment selection bias in an observational study, only subjects with complete information on potentially relevant confounding covariates were included. An \textit{a priori} list of 39 baseline covariates from over 100 available in the NTDB 5.0 was generated based on the literature and clinical suspicion as potential confounders of the relationship between treatment time and mortality and included the following: age, sex, race, maximum Abbreviated Injury Score (AIS)\textsuperscript{86} for each of the six anatomic regions (head/neck, face, chest, abdomen, extremity/pelvis, and skin), New Injury Severity Score \textsuperscript{19,21}, Glasgow Coma Score on arrival, first systolic blood pressure, blunt versus penetrating trauma, open versus closed femur fracture, number of femur fractures, number of serious extremity/pelvic injuries (AIS > 3), alcohol or drugs present at the time of admission, year of incident, time of arrival at hospital (day divided into four quarters), modified Charlson co-morbidity index\textsuperscript{87} and the presence or history of specific co-morbidities at admission (coronary artery disease including
heart failure; chronic obstructive pulmonary disease; stroke; hematologic disease including coagulopathy; diabetes; pregnancy; or hepato-renal disease or failure). Hospital characteristics include mean number of femoral shaft fractures admitted to the treating facility per year, teaching status, public versus private, American College of Surgeons trauma center level, and geographic region of the United States (Northeast, Midwest, Southern, Western). A subset of these covariates \( (W_m) \) was then identified that was associated with mortality \( (p < 0.2) \) using appropriate non-parametric bivariate tests of association adjusted for multiple testing with the Benjamini—Hochberg procedure. Observations with missing values in \( W_m \) were excluded in order to obtain a final sample with complete data in all measured potentially confounding covariates for analysis.

**Treatment Variables**

This study asks whether the timing of definitive treatment of femoral shaft fracture affects mortality. In order to provide clinically relevant categories for treatment beyond a simple dichotomy at 12 or 24 hours, 5 time periods were chosen \( a \ priori \), based on commonly used cut-points from the literature: \( t_0 \leq 12 \) hours, \( 12 < t_1 \leq 24 \) hours, \( 24 < t_2 \leq 48 \) hours, \( 48 < t_3 \leq 120 \) hours, \( t_4 > 120 \) hours. While 24 hours was the most commonly used threshold before which “early” treatment has been described in prior studies, I defined the referent group \( (t_0) \) as treatment within the first 12 hours in order to represent those subjects most likely to have been inadequately resuscitated and in a state of occult hypo-perfusion. I hypothesize that an additional physiologic stress from definitive fracture surgery in such patients could activate an adverse systemic response leading to end-organ injury, multi-organ failure and excess mortality when compared to those treated later when adequate resuscitation is more likely to have been achieved.

Each procedure in the dataset is recorded with a number of days, hours and minutes after admission when it was performed. Using this information, I calculated the time from admission to definitive fracture fixation, as defined above using ICD-9 procedure codes. By excluding patients transferred from other facilities or admitted one or more days after the injury, I assumed the calculated treatment time level was a reliable surrogate for time from injury, underestimating the true elapsed time only by the time required to reach hospital.

**Main Outcome Measure**

In-hospital mortality was the sole outcome. While other studies of treatment time effect have focused on Acute Respiratory Distress Syndrome or multi-organ failure, the reporting of such outcomes is inherently subjective and prone to measurement error. Moreover, one could not expect that such outcomes would be validly or reliably measured over such a large number or diverse range of treating centers. Other surrogate measures such as length of hospital stay are notoriously positively skewed, making analysis of the effect of timing difficult to analyze and interpret. In the absence of patient-centered outcomes (i.e. health-related quality of life) and systematic assessment of morbid events after treatment, mortality is the most objective, and therefore likely valid, endpoint. While potentially lacking sensitivity in detection of morbid consequences of treatment, mortality can be considered a very specific surrogate for clinically important morbidity.

**Statistical Methods**

The primary analytic strategy was based on marginal structural models (MSM) using the inverse probability of treatment-weighted (IPTW) estimator. This method uses the reciprocal
of the conditional probability of a subject receiving an assigned treatment given other covariates, as a means of confounding control in order to determine the effect of treatment. Causal effects are defined within the counterfactual 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 12 hours $P[Y(1)=1]$, to the risk corresponding to the hypothetical situation that each patient has definitive fixation within 12 hours $P[Y(0)=1]$. In the setting of a perfect experiment where patients could be perfectly randomized to treatment before or after 12 hours from admission, a possible estimate of the causal effect of interest is given by the observed relative risk of mortality, $P[Y=1|A=1]/P[Y=1|A=0]$. Since the relationship between the timing of surgery and subsequent clinical outcomes is confounded by baseline covariates such as age, injury severity or co-morbidities, the patients in our sample who were operated on after more than 12 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 12 hours.

Robins introduced a class of estimators that address this problem through a straightforward weighting approach. Like propensity score methods, 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 her observed treatment ($A_i$) given 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 pseudo-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 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 12 hours):

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

However, because of the complexity of clinically interpreting interaction by multi-level assigned treatment assignment, loss of statistical efficiency, and in order to test the hypothesis that delaying treatment beyond 12 hours was the main threshold beyond which the benefits in delay...
including adequate resuscitation are realized, a second set of weights were estimated using a model of binary treatment assignment dichotomized at 12 hours. Therefore, following conditional saturated weighted logit model with a treatment by covariate product term was used for assessment of interaction:

\[
\text{logit}(P(Y(a)=1)) = B_0 + B_1 \times I_{(a=t1-t4)} + B_2 \times V_j + B_3 \times I_{(a=t1-t4)} \times V_j
\]

Unlike propensity score analysis\textsuperscript{67, 72}, IPTW can easily be adapted for use with multi-level categorical treatment assignment. The likelihood of treatment within one of the five time categories is estimated using data-adaptive model selection that allows for complex functions of the variables to predict treatment (via splines and their multiplicative interactions)\textsuperscript{94} and goodness-of-fit of the model was assessed using the Hosmer-Le Cessie test\textsuperscript{95}. IPTW estimators only give consistent estimates of the parameters of interest if the treatment mechanism itself is estimated consistently. Since a mis-specified 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, assuming an \textit{a priori} functional form was avoided and instead a data-adaptive model selection algorithm was employed that chooses the functional form based on the information that is available in the data at hand. For this purpose, a model selection approach that is based on flexible polynomial spline functions was used and includes testing of all one-way interaction terms between candidate covariates\textsuperscript{94}. 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\textsuperscript{96}. Based on this observation, the selected model was complemented 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 Table 3.1 and Table 3.2 for the entire population and interaction assessments, respectively.

In order to contrast IPTW estimates with a conventional method of covariate adjustment, a logistic regression model was used to estimate the mortality risk as a function of timing of definitive fixation (coded as five indicator variables with \(t_0\) as the referent) and baseline covariates. Because, marginal structural models give population level (marginal) estimates and conditional models (such as logistic regression) provide pooled (conditional when covariates are included) estimates, direct comparisons are not appropriate. Nevertheless, because investigators often ascribe marginal inferences to these conditional estimates, this comparison was made for heuristic purposes. A backwards deletion approach was used to arrive at a parsimonious description of these relationships starting with a full model that contained main-effect terms for the treatment variable as well as \(W_m\). The baseline covariate with the largest p-value was then eliminated and the model refitted. Since the focus of this analysis was on estimating the effect of time until fixation on mortality, treatment time indicators were not considered for deletion from the model. This step was then repeated until all of the confounders had p-values smaller than 0.05. Model goodness-of-fit was assessed using the Hosmer-Le Cessie test\textsuperscript{95}.

Confidence intervals (CI) and p-values were non-parametrically obtained using a clustered bootstrap approach\textsuperscript{97, 98} that takes into account correlations between patients treated at the same hospital. 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 observation and would thus tend to overestimate the information available in the data, resulting in confidence intervals that are too small. Therefore, confidence intervals were estimated using the following “clustered” bootstrap approach. During each bootstrap iteration, a sample of size N was drawn with replacement from the pool of N hospitals in the 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 drawing samples of size n with replacement from the pool of n patients in the dataset. The same modified bootstrap approach was followed to obtain p-values by applying the general resampling-based methodology developed by Pollard and van der Laan. Results were considered to be statistically significant at p<0.05. Interaction was assessed by obtaining bootstrap-based p-values for the hypothesis that the relative risk of mortality associated with treatment time was identical for both levels (not serious and serious) defined by each of the four potential effect modifying associated injuries discussed above. Because interactions of this kind are difficult to identify, the level at which statistical significance is considered was relaxed to p<0.2. All analyses were carried out in R version 2.3.1. Data-adaptive model selection based on polynomial splines was performed using the polyspline package; multinomial regression models were fit using the multinom() function in the nnet package; the Hosmer-Le Cessie test was carried out using the resid() function of the Design package.

The results and testing of implicit assumptions of these methods are presented in Chapters 4 and 5.
<table>
<thead>
<tr>
<th>Covariate</th>
<th>12-24hrs</th>
<th>24-48hrs</th>
<th>48-120hrs</th>
<th>&gt;120hrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Injury Severity Score</td>
<td>1.00</td>
<td>1.01</td>
<td>1.03</td>
<td>1.04</td>
</tr>
<tr>
<td>Arrival between 6am and 12pm</td>
<td>0.23</td>
<td>0.77</td>
<td>0.70</td>
<td>0.94</td>
</tr>
<tr>
<td>Cardiac Co-morbidity</td>
<td>1.34</td>
<td>1.70</td>
<td>2.49</td>
<td>2.14</td>
</tr>
<tr>
<td>Number of serious associated extremity injuries</td>
<td>0.69</td>
<td>0.80</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>Maximum AIS Head Region</td>
<td>1.02</td>
<td>1.02</td>
<td>1.03</td>
<td>0.98</td>
</tr>
<tr>
<td>Bilateral Femoral Fractures</td>
<td>0.29</td>
<td>0.77</td>
<td>0.57</td>
<td>0.81</td>
</tr>
<tr>
<td>Age</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>Teaching Hospital</td>
<td>1.53</td>
<td>0.91</td>
<td>0.65</td>
<td>1.07</td>
</tr>
<tr>
<td>Glasgow Coma Score</td>
<td>1.00</td>
<td>0.96</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>Treated at Level 1 Trauma Center</td>
<td>0.95</td>
<td>1.36</td>
<td>1.42</td>
<td>1.40</td>
</tr>
<tr>
<td>Cerebro-vascular Co-morbidity</td>
<td>1.24</td>
<td>2.51</td>
<td>1.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Hospitals from Northeast Region</td>
<td>1.23</td>
<td>1.15</td>
<td>1.76</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Table 3.1: Summary of the estimates of the treatment mechanism that are used for the categorical mortality analysis. The entries in the first column give the factor by which the relative risk of being assigned to surgery between 12 and 24 hours rather than to surgery within 12 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-effects 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>
<thead>
<tr>
<th>Covariate</th>
<th>OR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrival between 6am and 12pm</td>
<td>0.52</td>
<td>(0.43, 0.63)</td>
</tr>
<tr>
<td>New Injury Severity Score</td>
<td>1.03</td>
<td>(1.02, 1.04)</td>
</tr>
<tr>
<td>Cardiac Co-morbidity</td>
<td>1.69</td>
<td>(1.37, 2.07)</td>
</tr>
<tr>
<td>Number of serious associated extremity injuries</td>
<td>0.77</td>
<td>(0.70, 0.84)</td>
</tr>
<tr>
<td>Age</td>
<td>1.01</td>
<td>(1.01, 1.02)</td>
</tr>
<tr>
<td>Glasgow Coma Score</td>
<td>0.97</td>
<td>(0.95, 0.99)</td>
</tr>
<tr>
<td>New Injury Severity Score (knot 50)</td>
<td>0.95</td>
<td>(0.92, 0.98)</td>
</tr>
<tr>
<td>Maximum AIS Head Region</td>
<td>1.01</td>
<td>(0.96, 1.07)</td>
</tr>
<tr>
<td>Bilateral Femoral Fractures</td>
<td>0.59</td>
<td>(0.30, 1.14)</td>
</tr>
<tr>
<td>Teaching Hospital</td>
<td>1.07</td>
<td>(0.91, 1.25)</td>
</tr>
<tr>
<td>Treated at Level 1 Trauma Center</td>
<td>1.18</td>
<td>(1.01, 1.39)</td>
</tr>
<tr>
<td>Cerebro-vascular Co-morbidity</td>
<td>1.54</td>
<td>(0.33, 7.24)</td>
</tr>
<tr>
<td>Hospitals from Northeast Region</td>
<td>1.25</td>
<td>(0.95, 1.64)</td>
</tr>
</tbody>
</table>

Table 3.2: Summary of the estimated treatment mechanism that is used for the binary mortality analysis (investigation of interaction). The entries give the estimated odds ratio for being assigned to surgery after 12 hours versus surgery during the reference time frame.

* Apart from main-effect terms for those covariates which are 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. Hosmer – Le Cessie test showed acceptable goodness-of-fit (p=0.54).
Chapter 4: Testing the Experimental Treatment Assignment Assumption

There are three basic assumptions required to estimate parameters for any form of marginal structural model (inverse probability of treatment-weighted, G-computation, or double robust)\textsuperscript{101}. The first is the *sequential randomization* assumption requiring that treatment be randomized with respect to the set of all possible counterfactual outcomes, given all prior covariates including treatments (as for time-dependent or repeated treatment regimes). This necessitates no unmeasured confounding be present. This assumptions is similar to Rosenbaum and Rubin’s “strongly ignorable treatment assignment” assumption for use of the propensity score\textsuperscript{67}. 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 is the assumption of *time ordering* essentially requiring that treatment precede outcome, which is safely preserved in this “point-treatment” study of treatment time effect on mortality. The third is the *consistency assumption* that bases valid inferences on the existence of counterfactual variables representing the set of subject specific outcomes had the subject, contrary to fact, had a treatment contrary to the one observed. This implies that the observed data are a subset of the full data; the rest can be considered missing.

Valid estimation of causal effects using inverse probability of treatment-weighted (IPTW) estimation, or any of the estimators just mentioned, obliges an additional *experimental treatment assignment* (ETA) assumption (this has also been called the positivity assumption)\textsuperscript{102, 103}. 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 injury scores of 3 or more are always operated on after more than 12 hours, none of the patients in the re-weighted sample who were operated on within 12 hours will have head 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 injury) or more characteristics. The ETA assumption is met in theory by study design: there were no conditions under which the probability of a subjects treatment time assignment within any of the five groups was set to 0 or 1. However, a situation where a certain baseline characteristic (W) or combination of W’s empirically drive the probability of treatment to (or near) 0 or 1 would constitute a practical violation of the ETA assumptions. Such practical violations of the ETA were investigated as follows. First, I performed a qualitative assessment by plotting observed treatment assignment against predicted values to assess whether for any value of the covariates (W\textsubscript{m}), the probability of treatment assignment is 0 or 1. Second, I present results from a quantitative method of checking ETA using Monte Carlo simulation. Finally, I present an empiric method for assessing the sensitivity of IPTW estimates to the most likely violation of ETA, that those treated after 12 hours from admission could not feasibly have been treated earlier. This approach is based on the standardized risk ratio (SRR) using modified IPTW weights that, in this case, gives an estimate of the effect of delaying surgery for those subjects who were treated early. Each of these methods assesses the validity of the marginal IPTW estimates by testing, either quantitatively or qualitatively, whether subjects could have been treated both early and late.
Graphical Evaluation of the Experimental Treatment Assignment Assumption

Mortimer\textsuperscript{104} introduced a method of graphically evaluating whether the ETA assumption is met practically, after the treatment model-fitting step has been undertaken. Using the treatment model for binary treatment dichotomized at 12 hours from admission introduced in Chapter 3 (Table 3.2), coefficients were used to calculate predicted treatment assignment for each subject. The log odds of treatment (based on coefficients derived from the treatment model) were plotted against both the observed and predicted treatment probabilities. Violations of the ETA assumption would be suggested by a probability of either treatment assignment equaling 0 or 1 for any value of covariates in the treatment model. Figure 4.1 shows this plot. It can be seen that for nearly all values of the covariates in the treatment model, treatment at both less than and greater than 12 hours occurred, and in all but one subject, the expected probability of treatment never fell below 5% or exceeded 95%. This visual representation supports the ETA assumption.

Monte Carlo Simulation Approach

Wang\textsuperscript{105} proposed the following simulation-based approach (parametric bootstrap) 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 \(W\). The marginal distribution of \(W\) is estimated by its empirical distribution using the data-adaptive approach described in Chapter 3 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 logistic regression. This model includes main-effect terms for \(A\) as well as all baseline covariates that were found to have significant bivariate associations with the outcome. These estimates now allow one to simulate realizations of the observed data structures by using the following sequential approach. First, \(n\) realizations of \(W\) are generated by sampling with replacement from the \(n\) observed values in the data set. Next, \(n\) realizations of \(A\) are drawn from the estimated conditional distribution of \(A\) given these simulated values of \(W\). Lastly, \(n\) realizations of \(Y\) are obtained 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 12 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 12 hours. The desired counterfactual mortality can then be estimated by simply taking the mean of these simulated \(Y\)-values.
Table 4.1 summarizes the results of such a simulation study for determining the extent of bias that our IPTW estimators of the marginal counterfactual mortality risks might be subject to due to a violation of the ETA assumption. Figure 4.2 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. Since all subjects seem to have adequately large probabilities of following any one of the five categories of treatment time 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.

**Standardized Risk Ratio Approach**

Standardization deals with the issue of confounding through estimating expected numbers of events and comparing them to those observed based on chosen weights and rates. Comparisons may be made between populations or treatment groups that have causal interpretations and inferences depend on the designated target population. Modified IPTW weights were proposed by Sato in order to expand causal contrasts afforded by IPTW estimates to specific components of the larger study group. Regardless of 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 treated 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 treated early. All components of the analysis described in Chapter 3 were identical for the 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 using 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.

\[ w_i = \frac{P(A=t_0|W=W_i)}{P(A=A_i|W=W_i)} \]

By using the early treatment group (t_0) as the standard population and the mortality experience of the entire study sample, the SRR is interpreted as the estimated risk ratio if those who were treated early where to have been treated at one of the later treatment-time periods. If subjects treated after 12 hours vary enough from those treated earlier to bias estimates of effect, the SRR—with its target population being the t_0 group—should differ from the marginal IPTW estimate. Similarities between SRR and IPTW would support exchangeability of those in the early treatment group with those treated later and therefore, support the ETA assumptions and the IPTW population estimate. I present results of this approach based on standardized risk ratios side-by-side with the IPTW results in Chapter 5 for direct comparison.
Chapter 4 Figure Legend

Figure 4.1: Plot of the expected probability ($Pr(\text{early12})$) and observed (early12) treatment assignment versus the expected log odds of treatment assignment (logoddsA). The dashed line represents 0.05 and 0.95 reference lines.

Figure 4.2: The distributions of IPTW estimates relative to the true parameter value obtained by G-computation (vertical red line).
Figure 4.1
Figure 4.2
<table>
<thead>
<tr>
<th></th>
<th>&lt;12hrs</th>
<th>12-24hrs</th>
<th>24-48hrs</th>
<th>48-120hrs</th>
<th>&gt;120hrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-computation truth</td>
<td>4.15%</td>
<td>2.27%</td>
<td>3.55%</td>
<td>2.68%</td>
<td>3.03%</td>
</tr>
<tr>
<td>Mean IPTW estimate</td>
<td>4.16%</td>
<td>2.20%</td>
<td>3.53%</td>
<td>2.64%</td>
<td>2.92%</td>
</tr>
<tr>
<td>Estimated bias</td>
<td>0.01%</td>
<td>-0.07%</td>
<td>-0.01%</td>
<td>-0.03%</td>
<td>-0.11%</td>
</tr>
</tbody>
</table>

Table 4.1: Estimate of ETA bias based on Monte-Carlo Simulation
Chapter 5: Results

The NTDB 5.0 included 18,404 subjects with at least one femoral shaft fracture who underwent internal fixation. Figure 5.1 illustrates how the study sample of 3069 subjects was selected based on inclusion and exclusion criteria. The average age was 32.7 years (SD 14.7) and there were 2,207 men (71.9%). More than half (N=1759) of the cases were treated within 12 hours of admission with decreasing numbers treated in each subsequent time period (Table 5.1). There were a total of 108 (3.52%) deaths in the study sample and the unadjusted estimate of the mortality risk was lowest in the group of patients treated between 12 and 24 hours (1.9%, 95% CI 0.7 – 3.3%) and increased for those treated later. Glasgow Coma Scale (GCS), New Injury Severity Score (NISS), and maximum AIS in the head and neck region were strongly associated (p< 0.0005) with both mortality and treatment, with a consistent trend towards more severe injury among patients treated later. 1,146 subjects (37.3%) had associated serious head or neck injuries, 1,696 (55.3%) had associated chest injuries, 808 (26.3%) had associated serious abdominal injuries, and 1,222 (39.8%) had an additional serious extremity or pelvis injury other then the femoral shaft fracture that was treated (Table 5.2). The most common serious associated injuries (AIS ≥ 3) were closed C2 fracture, pulmonary contusion, lumbar vertebral fracture and inter-trochanteric fracture, respectively.

Inverse Probability of Treatment-Weighted Analysis

The estimated mortality risk if all subjects had been treated in any one of the five time periods is shown in Figure 5.2, controlling for measured confounders. Mortality was estimated to be highest if all subjects were treated within 12 hours of admission. If treatment was performed between 12 and 24 hours, 48 and 120 hours, or after 120 hours, mortality was estimated to be reduced to 43% to 58% of this reference mortality level (Table 5.3). A mortality risk of 83% that of the reference (t0) level was estimated for treatment within 24 to 48 hour time frame, though this finding was not statistically significant. With the exception of the last time frame, the Hosmer – Le Cessie test showed acceptable fit of the selected treatment model (p-values for goodness-of-fit: 0.24, 0.56, 0.24, and 0.028 respectively).

In order to examine interaction by the presence of serious associated injuries and because the categorical analysis showed similar mortality risk among the latter four treatment groups, treatment was dichotomized at 12 hours. The marginal relative risk of mortality for later treatment was 0.65 (95%CI: 0.45 to 1.04; p-value: 0.06). For this dichotomous comparison, the data-adaptively selected logistic regression model for the treatment mechanism again showed acceptable goodness-of-fit (Hosmer – Le Cessie test p-value: 0.54). Figure 5.3 shows the estimated marginal mortality within groups defined by the presence of serious head, chest, abdominal or additional limb or pelvis injury, controlling for measured confounders. In general, more severely injured patients appear to benefit more strongly from delaying surgery for at least 12 hours. Only when stratifying on abdominal injury, did I see sufficient statistical evidence (Table 5.4) to suggest that the marginal relative risks for the two groups are different (p-value for effect modification: 0.09). The estimated marginal mortality with delayed treatment of those with no or low severity abdominal injuries is 82% (p-value: 0.39) of what it would be had they all been treated early. This is in contrast to a reduction to 36% (p-value: 0.03) of the risk of the mortality with early treatment, had all subject with serious abdominal injuries undergone delayed
treatment. While there was a marked decrease in estimated mortality for subjects with serious head injuries undergoing delayed treatment (RR: 0.62; 95%CI: 0.4 – 0.99; p-value: 0.04), there was insufficient statistical evidence to show a difference in effect as compared to subjects with no or low severity head injuries. Therefore, with the exception of severity of abdominal injury, it does not appear that sub-classification by any of the other tested associated injuries identifies subjects who would experience a greater or lesser degree of mortality reduction from delayed treatment than that estimated for the entire sample.

**Logistic Regression Analysis**

Table 5.3 summarizes the results of the estimation of the association using arbitrarily chosen multivariate logistic regression analysis giving odds ratios associated with treatment at any given time point conditional on holding other variables in the final model constant. All of the point estimates for delayed treatment times trended towards mortality reduction in comparison to treatment in under 12 hours, with the risk being lowest between 12 and 24 hours (OR: 0.52, 95%CI 0.18 – 1.19), though none reached statistical significance. Age, NISS and GCS were significant predictors of mortality. Also, both the number of femurs fractured and the total number of severe extremity or pelvic fractures were found to have a significant association with mortality. Surprisingly, being treated in a hospital from the Northeast was associated with a significant eight-fold increase in mortality versus subjects from the South. The Hosmer – Le Cessie goodness-of-fit test for this final logistic regression model suggested poor fit to the data (p=0.003). As for the IPTW interaction analysis, I worked with a dichotomized version of treatment time and found a similar difference in odd ratios for mortality among patients with low and high severity abdominal injuries (p=0.16) and no significant difference among the other subgroups tested (Table 5.5).

**Standardized Risk Ratios Analysis**

The estimates from the multivariate standardized risk ratio analysis (Table 5.3) were very similar to IPTW estimates with only slightly decreased precision. This analysis estimates the effect of treatment delay for those patients actually treated early. This provides the most conservative estimate of treatment effect by targeting those who, given their other prognostic factors, could have most likely been treated early. Again, I worked with a dichotomized version of treatment time and found a similar difference in standardized risk ratios for mortality among patients with low and high severity abdominal injuries (p=0.20) and no significant difference among the other subgroups tested (Table 5.6).
Chapter 5 Figure Legend

Figure 5.1: Study sample inclusion flow chart form the 8National Trauma Data Bank version 5.0 (2000-2005). bISS=Injury Severity Score. cWm= Covariates associated with mortality (p<0.2). (Reprinted with permission, Journal of Bone and Joint Surgery-American Edition)\textsuperscript{84}

Figure 5.2: Inverse Probability of Treatment Weighted mortality estimates with 95% confidence intervals (lines). (*) represents statistically significant relative risk reduction (p<0.05) versus the referent (t<12hrs) treatment group. (Reprinted with permission, Journal of Bone and Joint Surgery-American Edition)\textsuperscript{84}

Figure 5.3: Inverse-Probability-of-Treatment-Weighted mortality estimates stratified by severity of associated injury (presence of serious associated extremity/pelvis, abdominal, chest and head injury), with 95% confidence intervals (lines). (*) represents statistically significant relative risk reduction (p<0.05) versus the referent (t0) treatment group. AIS = Abbreviated Injury Score. (Reprinted with permission, Journal of Bone and Joint Surgery-American Edition)\textsuperscript{84}
Figure 5.1
Figure 5.2
Figure 5.3
<table>
<thead>
<tr>
<th>Covariates (W&lt;sub&gt;m&lt;/sub&gt;)</th>
<th>Association with treatment</th>
<th>Association with mortality</th>
<th>Treatment Groups Mean (SD)&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p-value</td>
<td>p-value</td>
<td>t&lt;sub&gt;0&lt;/sub&gt; (N=1759)</td>
</tr>
<tr>
<td>Glasgow Coma Score</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>12.67 (4.22)</td>
</tr>
<tr>
<td>New Injury Severity Score</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>27.35 (8.97)</td>
</tr>
<tr>
<td>Maximum AIS Head Region</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>1.71 (1.65)</td>
</tr>
<tr>
<td>Cardiac Co-morbidity</td>
<td>&lt;0.001</td>
<td>0.002</td>
<td>0.13 (0.34)</td>
</tr>
<tr>
<td>Number of serious associated extremity injuries</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>1.65 (0.94)</td>
</tr>
<tr>
<td>Arrival between 6am and 12pm</td>
<td>0.001</td>
<td>0.04</td>
<td>0.24 (0.43)</td>
</tr>
<tr>
<td>Age</td>
<td>0.01</td>
<td>0.01</td>
<td>31.86 (13.73)</td>
</tr>
<tr>
<td>Bilateral fractures</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02 (0.14)</td>
</tr>
<tr>
<td>Teaching Hospital</td>
<td>0.38</td>
<td>0.15</td>
<td>0.56 (0.50)</td>
</tr>
<tr>
<td>Hospitals from Northeast Region</td>
<td>0.54</td>
<td>0.00</td>
<td>0.08 (0.27)</td>
</tr>
<tr>
<td>Treated at Level 1 Trauma Center</td>
<td>0.69</td>
<td>0.05</td>
<td>0.39 (0.49)</td>
</tr>
<tr>
<td>Cerebro-vascular Co-morbidity</td>
<td>0.74</td>
<td>0.01</td>
<td>0.00 (0.04)</td>
</tr>
<tr>
<td>Mortality (95% Confidence Interval)</td>
<td>3.7% (2.3%, 5.4%)</td>
<td>1.9% (0.7%, 3.3%)</td>
<td>4.2% (2.2%, 6.4%)</td>
</tr>
</tbody>
</table>

Table 5.1: Summary table of bivariate associations of covariates with mortality and treatment group.

Treatment groups defined as: \( t_0 \leq 12 \text{ hours}, 12 < t_1 \leq 24 \text{ hours}, 24 < t_2 \leq 48 \text{ hours}, 48 < t_3 \leq 120 \text{ hours}, 120 \text{ hours} < t_4 \). *W<sub>m</sub>* = Subset of covariates with bivariate associations with mortality of \( p < 0.2 \). *SD* = Standard Deviation.
<table>
<thead>
<tr>
<th>Body Region of Associated Injury</th>
<th>Injury (ICD-9CM)</th>
<th>Frequency (number of diagnoses / total number of region specific AIS≥3 diagnoses)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head and Neck (N=1146: 202)</td>
<td>1. C2 Vertebral Fracture – Closed (805.02)</td>
<td>5.2% (84/1627)</td>
</tr>
<tr>
<td></td>
<td>2. Skull Base Fracture: Intracranial involvement and loss of consciousness – Not otherwise specified (801.00)</td>
<td>3.2% (52/1627)</td>
</tr>
<tr>
<td></td>
<td>3. Traumatic Subarachnoid Hemorrhage: Loss of consciousness – Not otherwise specified (852.00)</td>
<td>3.1% (51/1627)</td>
</tr>
<tr>
<td>Chest (N=1696: 51)</td>
<td>1. Lung Contusion – Closed (861.21)</td>
<td>40.1% (988/2447)</td>
</tr>
<tr>
<td></td>
<td>2. Traumatic Pneumothorax – Closed (860.0)</td>
<td>23.5% (574/2447)</td>
</tr>
<tr>
<td>Abdomen (N=808: 64)</td>
<td>3. Dorsal Vertebral fracture – Closed (805.2)</td>
<td>8.5% (207/2447)</td>
</tr>
<tr>
<td></td>
<td>1. Lumbar Vertebral Fracture – Closed (805.4)</td>
<td>38.4% (385/1004)</td>
</tr>
<tr>
<td></td>
<td>2. Splenic Parenchymal Laceration (865.03)</td>
<td>12.3% (123/1004)</td>
</tr>
<tr>
<td></td>
<td>3. Liver Laceration – Moderate (864.03)</td>
<td>10.5% (105/1004)</td>
</tr>
<tr>
<td>Extremity and Pelvis (N=1222: 126)</td>
<td>1. Inter-trochanteric fracture – Closed (820.21)</td>
<td>7.3% (127/1731)</td>
</tr>
<tr>
<td></td>
<td>2. Tibia with fibula fracture – Open (823.32)</td>
<td>6.1% (106/1731)</td>
</tr>
<tr>
<td></td>
<td>3. Pelvic fracture with disruption of pelvic ring – Closed (808.43)</td>
<td>4.9% (84/1731)</td>
</tr>
</tbody>
</table>

Table 5.2: Most common associated International Classification of Diseases, 9th Clinical Modification (ICD-9CM) coded injuries of the head and neck, chest, abdomen, and extremities and pelvis.

*AIS = Abbreviated Injury Score*
<table>
<thead>
<tr>
<th>Estimate</th>
<th>12-24hrs</th>
<th>24-48hrs</th>
<th>48-120hrs</th>
<th>&gt; 120hrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude Relative Risk</td>
<td>0.50 [0.06]</td>
<td>1.13 [0.69]</td>
<td>1.19 [0.52]</td>
<td>1.17 [0.77]</td>
</tr>
<tr>
<td>(0.18, 1.01)</td>
<td>(0.58, 2.05)</td>
<td>(0.67, 2.03)</td>
<td>(0.24, 2.65)</td>
<td></td>
</tr>
<tr>
<td>IPTW(^a) Relative Risk</td>
<td>0.45 [0.03]</td>
<td>0.83 [0.49]</td>
<td>0.58 [0.03]</td>
<td>0.43 [0.03]</td>
</tr>
<tr>
<td>(0.15, 0.98)</td>
<td>(0.43, 1.44)</td>
<td>(0.28, 0.93)</td>
<td>(0.10, 0.94)</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression(^b)</td>
<td>0.52 [0.15]</td>
<td>0.86 [0.67]</td>
<td>0.62 [0.14]</td>
<td>0.66 [0.43]</td>
</tr>
<tr>
<td>(0.18, 1.19)</td>
<td>(0.40, 1.68)</td>
<td>(0.31, 1.15)</td>
<td>(0.14, 1.55)</td>
<td></td>
</tr>
<tr>
<td>Standardized Risk Ratio(^c)</td>
<td>0.47 [0.07]</td>
<td>0.94 [0.85]</td>
<td>0.58 [0.09]</td>
<td>0.43 [0.05]</td>
</tr>
<tr>
<td>(0.14, 1.11)</td>
<td>(0.44, 1.76)</td>
<td>(0.21, 1.09)</td>
<td>(0.09, 0.94)</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3: Comparison of crude, regression, and marginal structural model estimates of effect of treatment time on mortality.

Point estimates are given with p-values in brackets [] and 95% confidence intervals in parentheses (). All analyses use \( t_0 (\leq 12 \text{ hours}) \) as the referent treatment group.  \(^a\)Inverse-probability-of-treatment-weighted (IPTW) population relative risk was derived using model for treatment assignment controlling for \( W_m \) (NISS, GCS, Northeast Region, age, arrival time, number of serious extremity/pelvis or head injuries, number of femur fractures, presence of cardiac or cerebro-vascular co-morbidities, teaching status, and ACS Level 1 designation).  \(^b\)Odds Ratios from final multivariate logistic regression model controlling for NISS, GCS, Northeast Region, age, arrival time, and number of serious (AIS \( \geq 3 \)) extremity/pelvis injuries.  \(^c\)Standardized Risk Ratio (SRR) using same treatment model as IPTW analysis but modified weights to give the estimated proportionate risk had those subjects treated early (\( t_0 \)) received treatment at a later time.
<table>
<thead>
<tr>
<th>Associated Injury</th>
<th>Low severity (AIS &lt;3)&lt;sup&gt;a&lt;/sup&gt;</th>
<th></th>
<th>High severity (AIS &gt; 3)</th>
<th></th>
<th></th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relative Risk (95% CI)</td>
<td>n</td>
<td>Relative Risk (95% CI)</td>
<td>n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extremity/ Pelvis Fractures</td>
<td>0.58 (0.33, 1.20)</td>
<td>1846</td>
<td>0.73 (0.39, 1.36)</td>
<td>1223</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>Abdomen</td>
<td>0.82 (0.54, 1.35)</td>
<td>2261</td>
<td>0.36 (0.13, 0.87)</td>
<td>808</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>Chest</td>
<td>0.67 (0.32, 1.37)</td>
<td>1373</td>
<td>0.64 (0.39, 1.11)</td>
<td>1696</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Head</td>
<td>0.71 (0.37, 1.77)</td>
<td>1923</td>
<td>0.62 (0.40, 0.99)</td>
<td>1146</td>
<td>0.74</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.4: Inverse-Probability-of-Treatment-Weighted estimates of the relative risk of mortality for treatment delay (greater than 12 hours) by severity of associated injury.

Reported p-values test the equality of treatment effects between low and high severity associated injury sub-groups. <sup>a</sup>AIS = Abbreviated Injury Score.
Table 5.5: Logistic Regression estimates of the relative odds of mortality for treatment delay (greater than 12 hours) by severity of associated injury.

Reported p-values test the equality of treatment effects between low and high severity associated injury sub-groups. *AIS = Abbreviated Injury Score.
Table 5.6: Standardized Risk Ratio estimates of the relative risk of mortality for treatment delay (greater than 12 hours) by severity of associated injury.

Reported p-values test the equality of treatment effects between low and high severity associated injury sub-groups. "AIS = Abbreviated Injury Score."
Chapter 6: Discussion

Clinical Comments

I have conducted an analysis of the effect of timing of definitive femoral shaft fracture fixation among multi-system trauma subjects in the National Trauma Data Bank. Using an inverse-probability-of-treatment-weighted (IPTW) analysis to estimate the marginal (i.e. the population level) risk of mortality for categorically defined treatment time, I found treatment in all but one (24 – 48 hours) of four delayed treatment categories to significantly lower risk of mortality to about 50% of that expected with early treatment (less than 12 hours). I also showed that patients with serious abdominal injuries have greater risk reductions from delayed fixation when compared to those with less serious or no abdominal injury. These findings were consistent with, but more precisely estimated than, the results of multivariate standardized risk ratios using the early treatment group as the standard population. These findings strongly support a cautious approach to early definitive femoral shaft fracture fixation among multi-system trauma patients, and especially those with serious associated abdominal injuries.

In two large studies assessing the effect of timing of internal fixation of femoral shaft fractures, subgroups of adult multi-system trauma patients (ISS>15) were defined with comparable demographics, associated injuries, and crude mortality rates. Fakhry found a significant decrease in mortality among later treated groups but only adjusted for ISS. Brundage did not report significant differences in mortality by treatment time but did find significant increases in complications such as pneumonia, ARDS, and length of hospital stay among patients treated between two and five days from injury. I divided the first hospital day into two periods, and found protective effects from delaying fixation even if only by twelve hours which may afford some degree of necessary resuscitation. Given that my sample was five times larger than the treated multi-system trauma subgroup in either of these studies, it is likely that increased power allowed more precise estimation of the protective effect of delayed treatment.

Several studies have asserted that patients with head or chest injuries may be more susceptible to the effects of early fracture fixation than patients without such associated trauma. Pape reported on a cohort of 106 multi-system trauma patients with femoral shaft fractures (ISS>18) with and without thoracic trauma (AIS>3). In a stratified analysis, these authors reported a significant increase in the incidence of ARDS among chest-injured patients who were treated with intra-medullary nailing early (<24 hours from admission) versus those who were not chest injured (33% versus 7.7%), and a similar, but not statistically significant, difference in mortality (21% versus 4%), suggesting that patients with chest injuries might benefit from delayed treatment. The notion that intra-medullary nailing potentiates the development of ARDS and further morbidity among patients who are vulnerable due to thoracic injury has been supported by human and animal studies showing embolization of marrow products during intra-medullary nailing, however the clinical importance of this phenomenon continues to be debated. Several observational studies using patients with thoracic injuries but without femoral shaft fractures as controls, have found no difference in incidence of ARDS or mortality suggesting that it is the severity of the pulmonary injury, rather than timing or mode of fixation, that determines the likelihood of adverse outcome. Among patients with associated head injuries, studies have shown hypotension and decreased cerebral perfusion pressure during early
fracture care \cite{28, 114, 115}. Still, increased mortality has not been shown in comparing early (<24 hours) to late treatment of femoral shaft fractures \cite{27, 29, 108, 116, 117}. While the relative risk for delayed treatment (beyond 12 hours) among those with serious thoracic injuries (RR 0.64%; 95%CI 0.39-1.11; p-value=0.10) approached significance and serious head injuries (RR 0.62%; 95%CI 0.40-0.99; p-value=0.04) achieved significance in my study, there was inadequate statistical evidence to support these subgroups experiencing a greater protective effect of delayed treatment than that expected in the rest of the sample.

The only factor for which there was evidence of modification of the effect of treatment timing on mortality was the presence of serious abdominal injury. For patients with serious abdominal injuries, undergoing delayed treatment had an estimated mortality risk 36% that of early treatment and this relative risk was significantly lower than for those without such injuries. Those with serious abdominal and extremity trauma may be particularly difficult to resuscitate because of significant on-going hemorrhage and potential for numerous vital organs to be injured. Several studies have shown abdominal injuries to be an important risk factor for mortality and morbidity in the multi-system trauma patient \cite{118, 119}. Schulman \cite{120} reported on a series of blunt trauma patients undergoing a standardized resuscitation protocol and found only ISS, abdominal injury (AIS $\geq 3$) and extremity and bony pelvis injury (AIS $\geq 3$) to be significantly associated with prolonged occult hypoperfusion. White \cite{119} reported a five-fold increase in the risk of ARDS when combined extremity and abdominal injuries were present. Patients treated early in my study with associated serious abdominal injuries were likely under-resuscitated and may have been particularly vulnerable to adverse outcomes from a definitive orthopedic procedure that often involves substantial additional blood loss.

Early treatment in my study is a likely surrogate for treatment in an under-resuscitated state were cellular hypoxia, oxidative stress, inflammatory response, and altered coagulation contribute to morbidity and mortality \cite{34, 35, 121, 122}. In a study of multi-system trauma patients with a femoral shaft fracture, Crowl \cite{30} showed that early intra-medullary fixation (<24 hours from admission) of femoral shaft fractures in the setting of occult hypoperfusion (serum lactate $>2.5$ mmol/L) leads to a significant increase in post-operative complications. Hypoperfusion resulting from trauma may prime the immune system for an inflammatory response if such treatment is undertaken prior to adequate resuscitations, and can lead to significant end organ injury \cite{123, 124}. The realization of this phenomenon has led to the description of so-called “damage-control orthopaedics” whereby definitive treatment is delayed until resuscitation of the patient has been adequately achieved \cite{25, 33, 125, 126}. My findings are consistent with this approach and signify the potential benefits that may be realized with delays in treatment of just 12 hours, particularly for those with associated abdominal injuries. The current end-points used to guide resuscitation, such as blood pressure, urine output, heart rate, base deficit and serum lactate are global markers that may underestimate occult tissue hypo-perfusion. In the future, more sensitive measures of tissue oxygenation measured by polarographic or near-infrared technologies \cite{127} and markers of inflammation and coagulation \cite{26, 128} that better reveal the physiologic condition of a patient are likely to replace simple temporal distinctions. They will provide more accurate criteria by which to determine when a patient is “ready” for definitive internal fixation of a femoral shaft fracture or other major orthopaedic injuries.
Methodological Comments

Marginal structural models with IPTW estimation were used in order to provide population level marginal estimates of effect with inference closest to what could have been achieved in the ideal setting of a randomized trial. Marginal structural models are based on the concept of counterfactual outcomes as introduced in Chapter 2. In the clinical scenario posed here, a subject’s counterfactual (or potential outcome) represents the set of outcomes had the subject, contrary to fact, had an exposure history other than the one actually observed. While this may appear to be an abstract concept, inferences provided are analogous to classical standardization of risk ratios. Moreover, several of the assumptions of marginal structural models such as no unmeasured confounding (sequential randomization) and temporal ordering are no different from those posed by conventional regression-based techniques of analysis of observational studies.

In Chapter 4, I introduced the assumptions necessary for valid identification and estimation of marginal structural model parameters generally, and those required for the use of IPTW specifically. Two of these assumptions are particularly important to note. 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. As discussed earlier, this assumption is ubiquitous in observational studies and cannot be empirically tested. While great care was taken to assess all possible confounders in a rich set of covariate data provided in the NTDB, we cannot be certain that some residual confounding is not present. The second assumption, called the experimental treatment assignment (ETA) assumption, requires that there are no values of the baseline covariates for which treatment is assigned in a deterministic fashion. The ETA assumption gets at the very heart of treatment selection bias that plagues therapeutic observational studies and those of surgical interventions specifically. The disparity between crude and adjusted estimates (Table 5.3), in light of higher injury severity and mortality in later treatment groups (Table 5.1), demonstrates the degree to which confounding by disease severity may threaten valid inferences in studies similar to ours. Because there is such a strong clinical tendency to prescribe a treatment based on the subject’s likelihood of outcome and danger for this to bias estimates of causal effects, I dedicated Chapter 4 to testing this assumption.

After applying stringent inclusion and exclusion criteria to selecting the study sample from the NTDB, three methods were undertaken to assess the ETA assumption. The first was qualitative evaluation using a graphical summary that showed that for nearly all values of covariate in the binary treatment assignment model, both treatment before and after 12 hours occurred. The second was an empirical, subject-matter-based approach based on the standardized risk ratio modification of IPTW proposed by Sato. Because subjects treated after 12 hours tended to have slightly higher injury severity (Table 1), one might postulate that regardless of the extensive confounding control performed here, unknown or unmeasured factors may have still preclude some patients from the later treatment groups from receiving early treatment and thus constituting a practical ETA violation. The most conservative response to this would be to assume that only those patients treated early could have reasonably been treated in any of the five treatment groups (t0-t4) and that a randomized trial including only those patients treated early might give different results compared to the marginal IPTW estimate. Using the early treatment group weights and population mortality rates to estimate the
proportionate change in mortality had those treated early been treated in each of the later treatment groups (Table 5.3), estimates of effect were comparable to IPTW estimates supporting the notion of sufficient exchangeability between the early treatment group and the larger study sample. Hence, the most conservative counterfactual scenario examined using multivariable standardized risk ratios supports the inference from the population-level IPTW estimate. These results are also in agreement with the Monte-Carlo simulations-based approach that examined to what extent the estimates might be biased due to such a violation of the ETA assumption. In all cases, the bias due to a possible violation of the ETA assumption appears to be minimal since all subjects seem to have adequately large probabilities of following any one of the five categories of time until definitive fixation.

In the absence of ETA violation, there are several additional advantages to IPTW over conventional regression analysis evident. Similar to a standardized risk ratio, IPTW provide unconditional estimates that are valid even when there is effect-measure modification. Comparing crude (unadjusted) estimates to estimates from both standard regression and IPTW shows evidence of confounding (Table 5.3). Apart from an initial relative risk reduction in the 12-24 hour range of about 50% that is consistent among all estimates, the crude estimates show a slight (though not statistically significant) increase in risk for those treated later. Both regression and IPTW show similar relative risk reduction in all groups treated after 12 hours, with IPTW estimates of effect being greater in magnitude and precision. There are several potential reasons for this. One is that as mentioned earlier, the two analytical methods typically estimate parameters that have different interpretations, regression typically being used to estimate the association conditional within sub-groups, and IPTW used typically for so-called marginal associations. While regression can also be used for marginal estimation, the advantage of IPTW is that the treatment mechanism is estimated semi-parametrically. The other reason is that while the accuracy and precision of logistic regression depends on goodness-of-fit of the outcome model—which was poor—it is the goodness-of-fit of the treatment model—which was acceptable—that matters for valid IPTW estimates. However, the approach based on IPTW may be considered inefficient because model selection is done orthogonal to the parameter of interest that has nothing to do with the treatment mechanism. Recently developed regression techniques using machine learning and targeted maximum likelihood optimize the model selection parameter that is more highly related to the parameter of interest. Such techniques may afford significant gains in efficiency over the IPTW approach presented here and warrant further investigation. Still, I believe that IPTW provides estimates with interpretation that closely approximate the causal inference that clinicians seek in understanding the impact of treatment time.

Another important feature of this analysis is the assessment of potential interaction by selected associated injury patterns. Like the IPTW estimates (Table 5.4), both the regression (Table 5.5) and SRR approaches (Table 5.6) yielded evidence of interaction between those with and without severe associated abdominal injuries. The IPTW analysis yielded a slightly larger difference in effect though these results are again, not directly comparable to those from standard regression and have a different target population from the SRR approach. Just as discussed earlier for main effects estimates, assessing interaction with standard regression gives inference that is conditional on the effect modifier of interest as well as any other confounders included in the model whereas IPTW estimates are marginal within strata of the effect modifier of interest. Whereas stratification or regression adjustment are used for confounder control and detection of
interaction in conventional analytical methods, marginal structural models make confounding control (IPTW) distinct from methods for detection of effect modification (adding treatment covariate interaction terms to an MSM). The marginal structural model including serious abdominal injury reveals a significant difference in the effect of early versus delayed definitive surgery depending on whether the abdominal AIS was less than 3 or greater than or equal to three. Moreover, within strata of abdominal injury severity, the estimate of effect is marginal in the same sense as the analysis with only the main effects term. The SRR interaction analysis supports the findings of IPTW as representing the most conservative causal contrast within strata of associate injury by using the early treatment group as the standard population. As with the main effects analysis, a marginal assessment of interaction gives inferences similar to that expected from sub-group analyses from a randomized controlled trial.

Final Comments

This observational study assessing the marginal effect of timing of definitive fixation of femoral shaft fractures in multi-system trauma patients showed a significant estimated relative risk reduction of about 50% among subjects treated between 12-24 hours and greater than 48 hours. Patients with severe abdominal injuries (AIS ≥3) showed significantly greater relative risk reductions than those with low abdominal AIS scores signifying interaction by this set of associated injuries. The observational nature of this study precludes absolutely ascribing causal effects to timing of treatment because the assumption of no unmeasured confounding cannot be empirically tested. In addition, there was missing information on measured covariates for subjects who would have otherwise been eligible for inclusion in the study sample (Figure 5.1) imposing a potential selection bias. While the NTDB includes a very large sample of trauma centers from all over the United States, it is still a convenience sample. However, this is the largest cohort ever used for investigation of this study question, and the national representation of academic and non-academic centers of various levels provides far-reaching generalizability of these results. The NTDB data provided extensive covariate information that was used for comprehensive covariate adjustment and robust analytical methods were used to optimize model fit and properly estimate confidence intervals in the setting of potential correlations by treatment center. Statistically efficient analytic methods were employed to address this bias and these methods provided unconditional estimates of relative risk. In order to compare marginal estimates to those given by conventional regression, IPTW results have been given side-by-side with logistic regression estimates and differences in interpretation discussed. Assessments of the experimental treatment assignment assumption gave both qualitative and quantitative support to the marginal IPTW estimates. Thus, these results provide strong empirical evidence in support of a delayed or “damage-control” approach to definitive fixation of femoral-shaft fractures among multi-system trauma patients and particularly among those with serious abdominal injuries. Further prospective randomized studies are necessary to confirm these results.
REFERENCES:


71. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American journal of epidemiology*. Aug 1 2003;158(3):280-287.


83. van der Laan M, Hubbard AE, Jewell NP. Estimation of Treatment Effects in Randomized Trials with Noncompliance and a Dichotomous Outcome. *U.C. Berkeley*


