2016

Endpoints In Intensive Care Unit Based Randomized Clinical Trials

Michael Oscar Harhay
University of Pennsylvania, michael.harhay@gmail.com

Endpoints In Intensive Care Unit Based Randomized Clinical Trials

Abstract

With few exceptions, intensive care unit (ICU)-based randomized clinical trials (RCTs) have failed to demonstrate hypothesized treatment effects. Undoubtedly, some of these failures are attributable to interventions that truly do not provide hoped-for benefits. However, this dissertation pursues the thesis that many null findings represent "false negatives" that are due not to ineffective therapies but to flawed study designs or analytic approaches. We examine the design and statistical methods traditionally employed in ICU-based RCTs, and their potential impacts on the efficient measurement and interpretation of treatment effects. Paper one presents a systematic review of 146 contemporary ICU-based RCTs in which we find that most trials were underpowered to detect small but potentially important mortality differences between treatment arms. We also find that the majority of RCTs (73%) specified primary outcomes other than mortality, that trials employing nonmortal primary outcomes more frequently identified significant treatment effects, and that both mortal and nonmortal endpoints were heterogeneously defined, measured and analyzed across RCTs. Thus, papers two and three focus on nonmortal endpoints, using ICU length of stay (LOS) as a case study to evaluate how best to measure and analyze duration-based nonmortal endpoints. In paper two, we conduct a statistical simulation study, demonstrating that nonmortal endpoints are interlinked with and confounded by mortality, and that the manner in which investigators choose to account for deaths in LOS analyses may influence their conclusions. In paper three, we examine another potential source of error in LOS analyses, namely the measurement error attributable to the additional ICU time that patients commonly accrue after they are clinically ready for ICU discharge. Using simulated data informed by our own ICU-based RCT, we demonstrate that this "immutable time" (which cannot plausibly be altered by the interventions under study) combines with clinically necessary ICU time to produce overall LOS distributions that may either mask true treatment effects or suggest false treatment effects. Our work provides evidence of the potential benefits and pitfalls when employing nonmortal outcomes in ICU-based RCTs, and also identifies a clear need for standardized methods for defining and analyzing such outcomes.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Epidemiology & Biostatistics

First Advisor

Scott D. Halpern

Keywords

epidemiology

ENDPOINTS IN INTENSIVE CARE UNIT BASED RANDOMIZED CLINICAL TRIALS

Michael Oscar Harhay

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

____

Scott D. Halpern, MD, PhD
Associate Professor of Medicine, Epidemiology, and Medical Ethics and Health Policy

Graduate Group Chairperson

____

Nandita Mitra, PhD, Professor of Biostatistics

Dissertation Committee

Jason D. Christie, MD, MSCE, Robert M. Kotloff/Nancy Blumenthal Professor for Advanced Lung Disease, Professor of Medicine and Epidemiology

David J. Margolis, MD, PhD, Professor of Dermatology and Epidemiology

Sarah J. Ratcliffe, PhD, Professor of Biostatistics

Dylan S. Small, PhD, Professor of Statistics

To Meera, Kiran and Nilan.

And to my father who always made education a priority,

even telling me to just find a master's degree and enroll.

Well, I did, and here we are.

ACKNOWLEDGMENT

I feel immensely grateful for having had the luxury of completing my PhD while my wife, Meera, and I built our family and home. Our relationship has followed the PhD curriculum. I met her while I was interviewing for programs, we had our wedding a month before my qualification exams, and our two children were both born a month prior to my thesis proposal and the first submission of this thesis. While hugs from my little girl can make any simulation study joyful, spending this time with Meera has reminded me each day how fortunate I am. I am indebted to many for the series of opportunities and events that supported and led to the completion of this work in this setting with my family. There are a few particular individuals that I'd like to thank here who were uniquely helpful when my plans weren't clear and my skills were much more limited.

Larry Hass, Jeremy Teissére, and Jeffrey Rudski allowed me to tinker and craft my undergraduate education at Muhlenberg, which I believe ultimately led me to Penn. In my early weeks as a graduate student, Arthur Caplan was uniquely generous to me, providing recommendations and opportunities that have put me on the path to completing this degree. And good fortune certainly played a role when I knocked on Darren Taichman's office door in 2005. He gave me my first job, supported me through two master degrees, and remains a close friend. I can never say thank you enough to Arthur and Darren for helping me establish myself at Penn.

Darren and Arthur were mentors to Scott Halpern, who at that time was a pulmonary fellow. During a lunch in 2007, Scott patiently listened to my grand plans (that have since changed several times over) and offered his name and mentorship to help me reach these aims. Though his career has blossomed, he has enthusiastically continued in this capacity ever since, most recently serving as my PhD mentor. If I am ever viewed as a quality scientist, it is to Scott and the research setting he created for me that I must attribute my success.

I am grateful to Scott for many other reasons, more than I can enumerate here, but at the top of the list is the introduction to Sarah Ratcliffe. Razor sharp, boundlessly patient and always available, Sarah worked weekly (and often daily) with me to find new ways to answer the hard questions Scott laid out before me to solve. The consummate mentor, she taught me statistical theory on her notepad, while also helping me choose among index funds for my family. Working with her has been one of the greatest joys of my PhD.

Many other individuals have been generous teachers to me. Dylan Small mentored me through my MS in statistics as well as serving on my dissertation committee. His contributions to my work are widespread, and I hope we continue to work together for many years. Jason Christie and David Margolis, despite their many professional responsibilities and roles, have remained approachable scholars who kindly agreed to be members of my dissertation committee. Their ideas, input, and experience have always elevated my work and I am grateful to them for their time and insight. Many years ago in Geneva I met and worked with Piero Olliaro. While our research and locations have diverged, his influence over my career and life goals is omnipresent.

I also want to thank Daniel Polsky and Andrew Epstein for the opportunity to learn econometrics as an analyst on their projects from 2011-2012. They put up with me when I didn't know much, and in doing so, that period in my academic life was intellectually transformative. Though I have often tried to thank them, it is hard to let them know how much they provided me.

Finally, I have had the fortune of sharing my experiences of doctoral studies (generally over beer) with Justin Brown, Henry Bergquist, Sean McElligott, and Kevin Haynes. To these individuals and the many other fantastic colleagues that I have met and worked with along the way, thank you.

Research reported in this dissertation was supported by the National Heart, Lung, and Blood Institute (F31-HL127947) and the National Cancer Institute (R01-CA159932 awarded to Scott Halpern) of the National Institutes of Health (NIH). I am enormously grateful for having had the good fortune of continuous funding support during the past 4.5 years from the NIH. Nonetheless, I must inform readers that the content of this work is solely my responsibility and does not necessarily represent the official views of the NIH.

Philadelphia, 2016

ABSTRACT

ENDPOINTS IN INTENSIVE CARE UNIT BASED RANDOMIZED CLINICAL TRIALS

Michael Oscar Harhay

Scott David Halpern

With few exceptions, intensive care unit (ICU)-based randomized clinical trials (RCTs) have failed to demonstrate hypothesized treatment effects. Undoubtedly, some of these failures are attributable to interventions that truly do not provide hoped-for benefits. However, this dissertation pursues the thesis that many null findings represent “false negatives” that are due not to ineffective therapies but to flawed study designs or analytic approaches. We examine the design and statistical methods traditionally employed in ICU-based RCTs, and their potential impacts on the efficient measurement and interpretation of treatment effects. Paper one presents a systematic review of 146 contemporary ICU-based RCTs in which we find that most trials were underpowered to detect small but potentially important mortality differences between treatment arms. We also find that the majority of RCTs (73%) specified primary outcomes other than mortality, that trials employing nonmortal primary outcomes more frequently identified significant treatment effects, and that both mortal and nonmortal endpoints were heterogeneously defined, measured and analyzed across RCTs. Thus, papers two and three focus on nonmortal endpoints, using ICU length of stay (LOS) as a case study to evaluate how best to measure and analyze duration-based nonmortal endpoints. In paper two, we conduct a statistical simulation study, demonstrating that nonmortal endpoints are interlinked with and confounded by mortality, and that the manner in which investigators choose to account for deaths in LOS analyses may influence their conclusions. In paper three, we examine another potential source of error in LOS analyses, namely the measurement error attributable to the additional ICU time that patients commonly accrue after they are clinically ready for ICU discharge. Using simulated data informed by our own ICU-based RCT, we demonstrate that this “immutable time” (which cannot plausibly be altered by the interventions under study) combines with clinically necessary ICU time to produce overall LOS distributions that may either mask true treatment effects or suggest false treatment effects. Our work provides evidence of the potential benefits and pitfalls when employing nonmortal outcomes in ICU-based RCTs, and also identifies a clear need for standardized methods for defining and analyzing such outcomes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1. INTRODUCTION

The growing burden and cost of critical illness

The demand for critical care, both in the United States (US) and worldwide, is outpacing the discovery of interventions that can substantively improve outcomes in intensive care unit (ICU) settings. In the US, one third of hospitalizations among patients older than 65 includes an ICU stay (Milbrandt et al., 2008) and 20% of the US population dies in an ICU (Angus et al., 2004). The global demand for and provision of critical care will likely grow in future years, both because of aging populations where ICUs are prevalent as well as the expansion of critical care in lower-income settings (Cook & Rocker, 2014; Fleischmann et al., 2016; Murthy et al., 2015).

The processes that lead to critical illness are diverse, which presents a challenge for researchers seeking to study and compare interventions in homogeneous ICU patient populations. Two paradigmatic examples of critical illness commonly encountered in ICU settings are sepsis/septic shock (Singer et al., 2016) and the acute respiratory distress syndrome (ARDS) (Force et al., 2012), both of which can result from acute or chronic illness and may present in a variety of patient types (e.g., various ages, comorbid conditions). ICUs also provide life-saving care for patients who are critically ill after surgery or trauma, as well as for a heterogeneous mix of patients with other pathologies (e.g., cardiovascular disease, cancer).

Despite the diversity inherent in critical care settings, one unifying theme of critical illness is its high cost, both in financial and human terms. In the US, 1% of the gross domestic product is spent on the provision of critical care, exceeding $80 billion per year, and representing approximately 3% of all health care spending (N. A. Halpern et al., 2016; N. A. Halpern & Pastores, 2010). In addition to the financial burden, survivors of critical illness are often left with physical, cognitive and psychosocial deficits that impede long-term quality-of-life (QOL) (Adhikari et al., 2011; Bienvenu et al., 2012; Fan et al., 2014; Herridge et al., 2011; Iwashyna, 2010; Kress

& Hall, 2014; Mikkelsen et al., 2012; Spragg et al., 2010). There is also a growing awareness of the downstream impact of critical illness on patients' caregivers, family members, and friends, who are called upon to cope with the loss of their loved ones after witnessing their suffering, or to provide daily care to survivors who require prolonged periods of time to regain their independence (Azoulay et al., 2005; Cameron et al., 2016; C. E. Cox et al., 2009). For instance, roughly one third of decision-makers for critically ill patients develop post-traumatic stress disorder or complicated grief that lasts months to years (Azoulay et al., 2005; Wendler & Rid, 2011). Finally, there is the lasting impact on the critical care workforce (i.e., physicians and nurses) who must face death frequently and often report on the futility of the care they deliver, resulting in high rates of burnout (S. D. Halpern, 2011b; Hamric & Blackhall, 2007; Meltzer & Huckabay, 2004; Piers et al., 2011).

Given the burden and costs of critical illness, innovations that improve outcomes among the critically ill have the potential to make vast impacts. Indeed, there is an impetus to advance all aspects of critical care, from the delivery of cost-effective care to the improvement of patient-centered outcomes including quality-of-life. However, studies investigating promising therapies and clinical interventions in ICU settings have met with limited success, a topic that will be further illustrated in the second chapter of this dissertation.

Randomization inference

Randomized clinical trials (RCTs) are considered the "gold standard" for producing the experimental evidence required to assess the efficacy and safety of interventions. The pursuit of an RCT-generated evidence base pervades all medical disciplines, and more recently has extended into the social sciences. This is because randomization, or randomly allocating patients to a study arm, probabilistically balances study arms on all pretreatment or baseline factors (measured and unmeasured) and thereby is able to mitigate the likelihood of selection biases when conducting hypothesis tests that compare study arms on post-randomization outcomes. As a result, random assignment of an intervention supports an estimate of the treatment effect that is independent of the error term. The result is an unbiased estimate of the impact of an intervention.

The attractions and benefits of an RCT over observational data to assess a hypothesis have been written about widely and presented in various frameworks over the past several decades. The counterfactual framework is useful for describing why RCTs are so highly valued in developing theories of causal inference between exposures and outcomes (Hernán & Robins, 2016; Morgan & Winship, 2015). Specifically, let A denote an exposure to an intervention in the ICU (A=1 indicates exposure to an intervention and A=0 indicates no exposure to an intervention). Then, for a binary outcome Y (e.g., death at day 30), we say that the intervention (A) has a causal effect on Y if Probability(Pr)$[Y^{a=1}=1] \neq \Pr[Y^{a=0}=1]$ and the intervention has no causal effect on Y (the null hypothesis) if $\Pr[Y^{a=1}=1] = \Pr[Y^{a=0}=1]$. Indistinguishable notation would be applied to any ICU outcome of interest regardless of if it was a continuous (e.g., length of stay [LOS]), time-to-event (e.g., time-to-resolution of delirium) or count (e.g., days of infection per 1,000 patient days) distribution.

In observational research, there is a concern about known and unknown (or observed and unobserved) confounders. The theory behind randomization is that if it is executed correctly, the concern about confounding at baseline is removed. Without randomization prior to exposure, there is no guarantee that A, representing the intervention, is uncorrelated with the error term, ε *(i.e., A* may be endogenous). The result of endogenous correlation is a potentially biased, or incorrect estimate of the impact of an intervention.

In contrast, the randomized experiment is built upon the concept of exchangeability, which by design is not susceptible to endogeneity. As a result, execution of an RCT is the closest a scientist can get to producing an unbiased causal effect estimate. Specifically, the risk of an event at baseline under the potential treatment value *a* among the treated is equal to the risk under the potential treatment value *a* for the untreated $\Pr[Y^a=1|A=1] = \Pr[Y^a=1|A=0]$. Said differently, the conditional risk of an outcome is equal in all subsets defined by treatment status in the population. Therefore, the baseline risk is equal to the marginal risk under treatment value *a* in the whole population. In ideal settings the process of randomization or random allocation should result in counterfactual data that is missing completely at random (MCAR) for each subject, such that causal effects can be estimated statistically without bias.

In practice, various logistical, patient and post-randomization factors can erode the assurance of unbiased effect estimates. Indeed, the focus of Chapters 4 and 5 of this dissertation are on post-randomization factors that can bias randomization inference. Specifically, in Chapter 4 we assess the impact of informative censoring from mortality, and in Chapter 5 we assess measurement error resulting from within-hospital patient flow and how these post-randomization processes can bias treatment effect estimates and interpretation.

Experimental evidence for treating critical illness

Efforts at improving critical care outcomes have resulted in a long history of ICU-based RCTs that have been unable to demonstrate statistically significant improvements in patient outcomes through new interventions, protocols, therapies and staffing models in the ICU (Aberegg et al., 2010; Angus et al., 2010; Annane, 2009; Ospina-Tascon et al., 2008). Exceptions include studies that have shown the benefits of low (rather than high) tidal volumes for patients receiving mechanical ventilation (ARDSnet Investigators, 2000), of restrictive (rather than aggressive) blood transfusion practices (Hebert et al., 1999; Villanueva et al., 2013), and of light sedation that is frequently interrupted (rather than heavy sedation without protocol-driven interruptions) to maintain comfort among ventilated patients (Girard et al., 2008; Kress et al., 2000). Unfortunately, the vast majority of critical care RCTs have not demonstrated interventions that decreased mortality (Aberegg et al., 2010; Landoni et al., 2015; Ospina-Tascon et al., 2008). A review of RCTs published exclusively in the journal *Intensive Care Medicine* from 2000-2010 found an overall success rate of 48.8% (of 221 RCTs) (Latronico et al., 2013). However, in two reviews of RCTs where mortality was the primary endpoint, success rates were only 14% (10 of 72 RCTs published before August 2006) (Ospina-Tascon et al., 2008) and 18% (7 of 38 RCTs published from 1999-2009 in 5 major medical journals) (Aberegg et al., 2010), respectively. Given that these studies all focused on published RCTs, the true rate of positive RCTs is potentially lower as negative studies, especially industry-sponsored trials, may be less likely to be submitted for publication or ultimately be accepted for publication. The low rate of successful ICU-based RCTs has not gone unnoticed; there are some thought leaders who have been so disappointed by these trends that they have suggested entirely abandoning the concept of RCTs in the ICU (Dreyfuss, 2004; J.-L. Vincent, 2010). However, it is unclear whether the majority of ICU-based RCTs were negative because of a true lack of treatment effect or because of the design elements of the RCTs in which they were tested (J. L. Vincent, 2009). This is especially relevant in studies of nonmortal clinical endpoints (e.g., LOS in the ICU), where the statistical handling of dropout (censoring) from death could impact the interpretation of results. Therefore, this dissertation seeks to expand the empirical solutions available to researchers to help in distinguishing negative versus misinterpreted trials. To do so, we build on a small and limited empirical foundation.

While there is a range of proposed explanations (*see conceptual framework in Table 1.1*) for the low success rates of ICU-based RCTs, empirical research has focused almost exclusively on explanations related to statistical power in mortality studies. Specifically, researchers have identified a practice termed *delta inflation*, wherein unrealistically large predicted treatment effects are used to estimate a trial's needed sample size (Aberegg et al., 2010; Latronico et al., 2013). Conversely, detection of smaller (but possibly more realistic) mortality differences between study arms requires larger study samples. As a result, studies that are powered based on delta inflation may be perceived as inconclusive or negative because potentially clinically relevant treatment effects are not statistically significant. The empirical assessment of delta inflation bias has been restricted to RCTs of mortality. It is unclear if this practice occurs with other nonmortal primary outcomes, and further, if nonmortal endpoints are as frequently negative as studies powered to detect a treatment-associated decline in mortality. It is also unclear if misspecification of other elements of the power calculation, such as the event rate in the control arm of the study, leads to underestimated necessary sample sizes. These specific questions are examined in Chapter 2.

This thesis also pursues a specific focus on nonmortal outcomes which are largely under-scrutinized but increasingly advocated trial endpoints by investigators and trial consortiums (Mebazaa et al., 2016; Opal et al., 2014; Spragg et al., 2010; Young et al., 2012). Principal to this endeavor is the identification and subsequent standardization of core outcomes that will be measured and analyzed identically across future trials to promote less biased comparisons between different trials and promote harmonized data assemblage in meta-analyses (Blackwood et al., 2014; Blackwood et al., 2015). This area of research activity is very nascent in critical care, but has seen much activity in other disciplines through the COMET (Core Outcome Measures in Effectiveness Trials) Initiative which focuses on the development and application of a standardized set of outcomes (Prinsen et al., 2014; Williamson & Clarke, 2012). Among the several goals of this work, the research herein seeks to enhance research in critical care by considering standardized analytic methods to improve the detection of clinically relevant treatment effects and facilitate comparisons across ICU populations worldwide.

Dissertation aims

As reviewed above, RCTs among critically ill patients commonly fail to detect their hypothesized treatment effects, but it is unknown whether these trials have correctly identified the lack of treatment effect (i.e., true negative) or have committed a type-II error (i.e., false negative) due to methodologic flaws. Therefore, this dissertation seeks to evaluate the hypotheses outlined in Table 1.1. The overall objective is twofold: (1) to provide and advance knowledge that will improve current approaches to designing RCTs in critical care and (2) advance novel perspectives and concepts to improve the assessment of experimental evidence from ICU-based

RCTs. These two goals are collectively accomplished through a series of three thematically linked analyses that elucidate some of the potential mechanisms underlying the ongoing challenges that past trials have encountered in identifying treatment effects.

First, in Chapter 2, we conduct the largest-ever study of the outcomes, design, and analysis of ICU-based RCTs published in 16 leading journals from 2007-2013. In Chapter 3, we present the empirical framework for Chapters 4 and 5, which is based on the finding that the majority of RCTs studies in Chapter 2 were designed to assess a nonmortal primary endpoint. Therefore, Chapters 4 and 5 focus on issues salient to RCTs with nonmortal endpoints, and use ICU LOS as a case study. Indeed, LOS is the most frequently used nonmortal outcome across all published trials (Chapter 2), and both a patient-centered and critical operational outcome. It is also representative of a broader class of endpoints assessing “durations,” such as the duration of organ dysfunction, delirium, or ventilation. Accordingly, LOS is an illustrative endpoint to appraise the empirical and conceptual issues related to the definition, measurement and statistical comparisons of nonmortal measures between study arms.

First, in Chapter 4, a detailed examination of the epidemiological and statistical issues of measuring and analyzing ICU LOS in the presence of informative censoring due to mortality is undertaken. Then, in Chapter 5 we identify and evaluate the importance of a new form of measurement error termed ‘immutable time bias.’ This bias is defined as immutable because the extra time contributed to the total LOS *could not be altered* by the intervention, but rather is driven by system issues including floor bed availability, capacity strain, or administrative delays. With a simulation study informed by both the Study to Understand Nighttime Staffing Effectiveness in a Tertiary Care ICU (SUNSET-ICU) RCT, performed at our institution, and the few other RCTs we identified that reported the “ready-to-discharge time” over “actual discharge time,” we assess the identification of treatment effects in LOS under different hypothetical scenarios. We summarize the results of these three empirical investigations and their relevance for future ICU-based RCTs in Chapter 6.

Table 1.1. Hypotheses to explain low efficacy in critical care randomized clinical trials

Domain	Dissertation chapter assessing an element of this hypothesis	Hypothesis
Intervention	2	The proposed interventions are not truly effective interventions.
Logistical	2,5	RCTs are sufficiently powered but patient attrition leads to appreciable post-randomization losses so that the intention-to-treat analyses are highly conservative or biased.
Study population	2,4,5	Treatment-effect heterogeneity may lead to a diluted effect estimate because while interventions work for certain patients, others are too sick and/or have too many competing risks for death for singular interventions to be of benefit.
Power	2,4,5	RCTs may suffer design issues, such as insufficient power to detect relatively small but important effects in appropriate outcomes (i.e., excessive Type II error rates).
Outcome	2,4,5	Outcome measures are inappropriate, that is, the intervention does not impact the outcome that is measured or the selected outcome is not the ideal way of measuring an effect.
Analysis	2,4,5	Outcome measures themselves are appropriate, however, the mathematical methods of evaluating them are flawed or limited.

CHAPTER 2. OUTCOMES AND STATISTICAL POWER IN ADULT CRITICAL CARE RANDOMIZED TRIALS

This chapter has been published in the *American Journal of Respiratory and Critical Care Medicine*, official journal of the American Thoracic Society, and is reprinted here with permission of the American Thoracic Society. Copyright © 2014 American Thoracic Society. The citation for this publication is:

> Harhay MO, Wagner J, Ratcliffe SJ, Bronheim RS, Gopal A, Green S, Cooney E, Mikkelsen ME, Prasad Kerlin M, Small DS, Halpern SD. 2014. Outcomes and Statistical Power in Adult Critical Care Randomized Trials. *American Journal of Respiratory and Critical Care Medicine* Jun 15;189(12):1469-78.

Introduction

In this chapter we examine the design, analysis and outcomes used in published ICU-based RCTs. As noted in the introduction, the primary motivation for this analysis is that most published RCTs of critical care interventions that aim to reduce mortality have produced negative results (Aberegg et al., 2010; Angus et al., 2010; Annane, 2009; Ospina-Tascon et al., 2008), and even these reports may be overly optimistic because negative trials are less likely to be published and identified. While several RCTs have revolutionized critical care practice (ARDSnet Investigators, 2000; Girard et al., 2008; Guerin et al., 2013), the results of critical care trials on the whole have been so disappointing that some leaders in the field have suggested a renewed focus on non-experimental study designs (Dreyfuss, 2004; J.-L. Vincent, 2010).

However, truly negative trials are valuable because they prevent the use of interventions that are either costly but non-beneficial or even harmful (e.g., intensive insulin therapy (Van den Berghe et al., 2006) and hydroxyethyl starch (Myburgh et al., 2012; Perner et al., 2012)). Further, there are many reasons why trials may not demonstrate a treatment effect, including ineffective interventions, difficulty recruiting adequate sample sizes, post-randomization patient attrition, heterogeneous patient populations or treatment-effect heterogeneity, use of inappropriate outcomes, unreasonable assumptions (e.g, predicted effect sizes) used in power calculations and/or smaller than appreciated attributable morbidity and mortality fractions (Aberegg et al., 2010; Angus et al., 2010; Annane, 2009; Marini, 2006; McAuley et al., 2010; Ospina-Tascon et al., 2008; Reade & Angus, 2009; Rubenfeld & Abraham, 2008; van Meurs et al., 2008). Understanding an evidence base requires the ability to distinguish among these reasons so as to differentiate trials that are truly negative from those that may be falsely negative.

As a first step in enhancing understanding of clinical trials in adult critical care, we created a contemporary database of the design, analysis, and reporting of ICU-based RCTs. Herein, we describe the development of this database, the characteristics of RCTs published in the past 6 years with a specific focus on the outcome measures used, the quality of these RCTs using selected quality metrics, and the extents to which several issues germane to statistical power may contribute to trials' outcomes.

Methods

A group of physicians, epidemiologists and statisticians, guided by the 2007 CONSORT (Hopewell et al., 2008) (Consolidated Standards of Reporting Trials) statement, Jadad scale (Jadad et al., 1996; Juni et al., 2001), and prior work and commentaries on the topic (Aberegg et al., 2010; Angus et al., 2010; Annane, 2009; Chiche & Angus, 2008; Marini, 2006; McAuley et al., 2010; Ospina-Tascon et al., 2008; Reade & Angus, 2009; Rubenfeld & Abraham, 2008; van Meurs et al., 2008; J.-L. Vincent, 2010) identified RCT elements to be abstracted. We began our search for published RCTs in January 2007, as this approximated the end of prior review periods (Aberegg et al., 2010; Ospina-Tascon et al., 2008) through May 2013. We examined only RCTs of diagnostic, therapeutic, or process and systems interventions among adult patients conducted in an ICU published in 16 prominent general or critical care journals (Table 2.1).

Table 2.1. Eligible journals and published critical care randomized clinical trials abstracted

Peer-reviewed journal	*Number of RCTs*
Critical Care Medicine	44
Intensive Care Medicine	20
JAMA	17
New England Journal of Medicine	17
American Journal of Respiratory and Critical Care	14
The Lancet	10
Chest	7
Anesthesia and Analgesia	4
Anesthesiology	2
Annals of Internal Medicine	2
Archives of (now JAMA-) Internal Medicine	2
British Journal of Anesthesia	2
Canadian Medical Association Journal	2
British Medical Journal	1
Journal of Critical Care	1
Journal of Trauma and Acute Care Surgery	1

The MeSH terms “Intensive Care Unit”, “Critical Care”, “Critically Ill”, “Intensive Care”, “ICU”, “Randomized clinical trial”, and “Randomized controlled trial” were combined with the unique search names for the targeted journals. Search filters were used to limit our search to studies of adults that were published in the English language. The search results were screened for duplicates using RefWorks (ProQuest; Ann Arbor, MI) to create a single list of unique articles for eligibility screening.

For an RCT to be considered eligible it had to: (i) be published in one of the sixteen pre-specified journals no earlier than 2007, (ii) take place in one or more ICUs (i.e. not in an emergency department, post-anesthesia recovery unit, or elsewhere), (iii) include adult patients, and (iv) specify a primary clinical outcome (Table 2.2).

Table 2.2. A priori selected outcomes

1. ICU mortality
2. In-hospital mortality
3. 28-day mortality
4. 29-180 day mortality
5. 181+ day mortality
6. ICU readmission
7. Hospital discharge disposition
8. Costs/ Charges
9. ICU length of stay
10. Hospital length of stay
11. Ventilator-free days
12. Duration of mechanical ventilation
13. Organ failure-free days
14. Patient, family, physician, nurse, or other provider satisfaction
15. Complications/ adverse outcomes
16. Healthcare associated infections
17. Quality of life
18. Survival
19. Incidence of acute organ failure
20. Delirium
21. Composite or other outcome (not previously specified)

We excluded intermediate and physiologic outcomes because our goal was to identify trials testing interventions that were sufficiently mature as to be applied clinically as opposed to those that were primarily hypothesis generating. Physiological and psychological test scores were not considered to be clinical outcomes and hence were not abstracted. Intermediate and physiologic outcomes were excluded because our goal was to identify trials testing interventions that were sufficiently mature as to be applied clinically as opposed to those that were primarily hypothesis generating.

Data abstraction

Using the Research Electronic Data Capture (REDCap) platform hosted at the University of Pennsylvania (Harris et al., 2009), two investigators independently abstracted the primary and secondary outcomes, as reported by the authors in each trial, and the result (positive or negative) for each RCT. We relied on the data as reported by the original authors in their publication for each study during abstraction. Three investigators (RB, AG, SG) served as primary data abstractors, with two of them initially screening each article identified by the electronic search for initial eligibility. To validate this screening process and as an internal quality measure, four other investigators (MOH, JW, EC, and SDH) screened four of the sixteen selected journals over the full duration of the inclusion period (n=951 journal issues). This screening identified 24 RCTs, of which 18 were eligible per inclusion criteria. The 3 primary data abstractors achieved perfect agreement, identifying all 24 of these RCTs and correctly excluding the 6 ineligible trials.

A superiority study was considered positive if the p-value for the analysis of the primary outcome was less than 0.05, or the adjusted significance level after interim analyses, based on the reporting in each RCT. An equivalence or non-inferiority study was considered positive if the difference between study arms fell between the pre-determined margins (confidence intervals) and met the equivalence or non-inferiority hypothesis at the p-value declared by the study's authors. When a study had more than two arms, outcomes were recorded from the control arm and the arm employing an intervention of maximal dose or degree. Data were also extracted on study funding, type of intervention tested, target patient population, enrollment and retention, and statistical power.

To assess statistical power, we abstracted three specific methodological elements: (1) discussion of the power calculation used for the trial, (2) rationales for the parameters used in the sample size or power estimation, and (3) participant accrual. Discussion of the power calculation was defined as reporting the inputs used to calculate power or sample size, such as the baseline (control group) event rate and the expected treatment effect size for binary endpoints. The rationales for sample size or power estimation inputs could include prior research results, pilot studies, or other objective data. Participant accrual was tracked by assessing CONSORT diagrams, when available, indicating the number of patients screened, randomized and ultimately analyzed (Hopewell et al., 2008).

The data abstractors achieved greater than 90% agreement for individual data elements, including primary and secondary outcomes, funding, target sample size, and reason(s) for study exclusion. The first author adjudicated the discrepancies that arose. STATA 13 (StataCorp, College Station, Texas) was used for database management and analysis.

Analyses to assess statistical power

For each RCT with a binary outcome, we abstracted the predicted and observed risk difference on the absolute scale. We used the absolute, rather than relative, risk difference because absolute differences are used to determine the clinical significance of effects (Sackett et al., 1997). For example, to calculate the number needed to treat, the absolute risk reduction is required. For negative trials, we evaluated whether (non-significant) reductions in the primary outcome of 3% or greater were identified. Our choice of a 3% cutoff is somewhat arbitrary, but was chosen *a priori* based on the view that any treatment-associated absolute effect of this size would clearly be important to patients, and that effects less than 3%, albeit potentially important, could also more easily be attributable to noise or random error. These assessments were limited to trials that reported power calculations, so as to enable uniform determinations of whether or not these RCTs were powered to document these effect sizes as significant.

We also explored the related phenomenon of “delta inflation bias” (Aberegg et al., 2010; Latronico et al., 2013), whereby unrealistically large treatment effects are predicted in power calculations, resulting in target sample sizes that may fail to detect clinically important differences. To estimate the necessary sample size for a trial there are two essential components, the rate of the outcome in the control arm and the expected difference, termed the minimally clinically important difference (MCID). The MCID is also known as th**e** predicted treatment effect, effect size or delta (D). The MCID characterizes the smallest change in the primary outcome that is felt to be meaningful to both the clinician and patient. The MCID is the most variable and important component of sample size calculations, even when the outcome upon which the trial is powered is the same across studies. To properly design a RCT that can adequately answer the primary study question, it is necessary to establish the magnitude of the difference in the primary end point that will signify a clinically relevant treatment effect. Detection of a smaller difference between study groups requires larger sample sizes. Detection of a larger difference between study groups requires smaller sample sizes. This mathematical tension is the speculated motivation for researchers to select a larger threshold that consequently decreases the sample size requirements. This is delta inflation (Aberegg et al., 2010). The corollary of this practice is the increased likelihood of type II error (not detecting a true effect).

To understand how delta inflation works we can see below that if we seek to determine the targeted sample size (*n*) using significance level α, (typically 0.05) and want to have power $1-\beta$ (typically 0.80) when we an assumed value for P_1 (baseline mortality) & unknown value for p_2 (mortality in the treated arm), a speculated mortality decline must be presumed. This decline creates that delta, that is, $\Delta = P_1 - P_2$. These components permit the calculation of a targeted sample size based on that delta using the following equation that can be calculated in most statistical packages: $n_1 = \frac{\left[\sqrt{\overline{p}\overline{q}\left(1+\frac{1}{k}\right)}z_{1-a/2} + \sqrt{p_1 q_1 + \frac{p_2 q_2}{k}}z_{1-b}\right]^2}{\mathrm{D}^2}$ If a researcher knows that a certain sample size range will be available the delta can be manipulated to attain an acceptable power (e.g, 0.80). This can be seen in an alternative configuration of the above equation that can be used to solve for a certain power. For instance, if we must use specific sample sizes n_1 & n_2, the assumed values for P_1 & P_2, and hence $\Delta = P_1 - P_2$, can be changed and then the power we achieve is given by: $\Phi\left[\frac{\Delta}{\sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2}} - z_{1-\alpha/2}\frac{\sqrt{\overline{pq}(1/n_1 + 1/n_2)}}{\sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2}}\right]$ where $\overline{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$.

Using the actual enrolled sample sizes in the control and treatment arms, and the observed baseline mortality rate, we calculated the power of each study to observe a clinically significant, treatment-associated mortality reduction from of 3% to 15%. The number of trials with 80% power was tallied for each treatment-associated mortality reduction of 3% to 15%. We then re-did each calculation using the predicted baseline mortality from the published RCT. Comparison of power obtained using the observed and predicted rates therefore highlights how often mis-predictions of baseline event rates influence power. Similar analyses were undertaken using all RCTs with binary nonmortal primary outcomes.

Statistical analyses

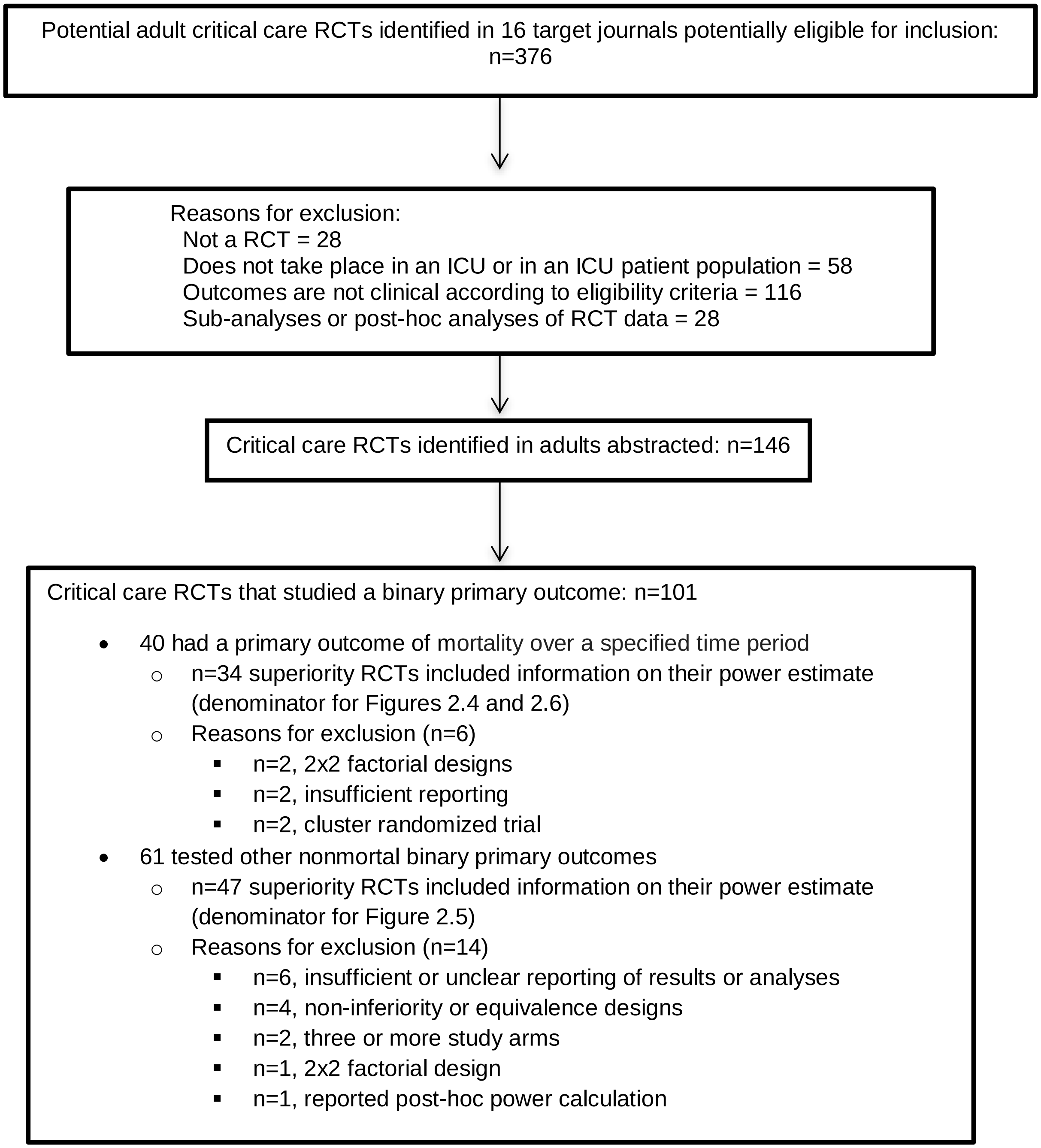
We conducted unadjusted comparisons of proportions using x^2 tests to examine the differences in proportions of successful trials across trial characteristics. We used multivariable regression to identify study-level characteristics associated with a trial's being positive. For this purpose, given limited degrees of freedom, we limited our assessments to the following trial characteristics: (1) mortal vs. nonmortal primary outcome, (2) funding source (3) single vs. 2-10 centers vs. > 10 centers and (4) type of intervention. Odds ratios (ORs) from a logistic regression and prevalence rate ratios (PRR) from a Poisson regression with a robust variance estimator are presented since the ORs will overestimate relative risks with event (positive trial) rates >10% (Deddens & Petersen, 2008).

Results

Our search identified 376 potential studies published between January 2007 and May 2013 (Figure 2.1). Of these, 146 met the pre-specified inclusion criteria. The most commonly tested types of interventions were protocols (49%) and drug therapies (40%) (Table 2.3). Most trials (92%) compared two intervention arms (max=5). Overall, 54 (37%) were positive, that is, these RCTs demonstrated a significant difference between study groups in the primary outcome as hypothesized (Table 2.3). In addition to the 21(14%) RCTs stopped early for safety or futility an additional 4 RCTs (3%) revealed statistically significant findings of inferiority (i.e., effects contrary to the primary hypothesis).

The most common primary outcomes were measures of mortality over a specified time period (27%), followed by outcomes related to healthcare associated infections (23%), ventilation (21%) (e.g., time to extubation, ventilator-free days or required mechanical ventilation), and quality (10%) (e.g., complications or adverse events). The incidence of positive trials varied depending on the primary outcome. The success rates for trials using these four above-mentioned outcomes were 10%, 58%, 43% and 50%, respectively. Two of the four positive mortality trials were only significant after pre-specified adjustment (Jansen et al., 2010; Papazian et al., 2010); thus, only 5% of these trials showed statistically significant differences in crude mortality rates. Twenty-four of the 40 trials where mortality was the primary outcome studied 28- or 30-day mortality. Five additional RCTs included a mortality endpoint as part of a composite primary outcome with nonmortal measures and one RCT was powered on mortality despite being listed as a secondary outcome; of these, one trial was positive.

Figure 2.1. Analytic sample of published randomized clinical trials of critical care interventions



Potential adult critical care RCTs identified in 16 target journals potentially eligible for inclusion: n=376

Reasons for exclusion:
Not a RCT = 28
Does not take place in an ICU or in an ICU patient population = 58
Outcomes are not clinical according to eligibility criteria = 116
Sub-analyses or post-hoc analyses of RCT data = 28

Critical care RCTs identified in adults abstracted: n=146

Critical care RCTs that studied a binary primary outcome: n=101

- 40 had a primary outcome of mortality over a specified time period
    - n=34 superiority RCTs included information on their power estimate (denominator for Figures 2.4 and 2.6)
    - Reasons for exclusion (n=6)
        - n=2, 2x2 factorial designs
        - n=2, insufficient reporting
        - n=2, cluster randomized trial
- 61 tested other nonmortal binary primary outcomes
    - n=47 superiority RCTs included information on their power estimate (denominator for Figure 2.5)
    - Reasons for exclusion (n=14)
        - n=6, insufficient or unclear reporting of results or analyses
        - n=4, non-inferiority or equivalence designs
        - n=2, three or more study arms
        - n=1, 2x2 factorial design
        - n=1, reported post-hoc power calculation

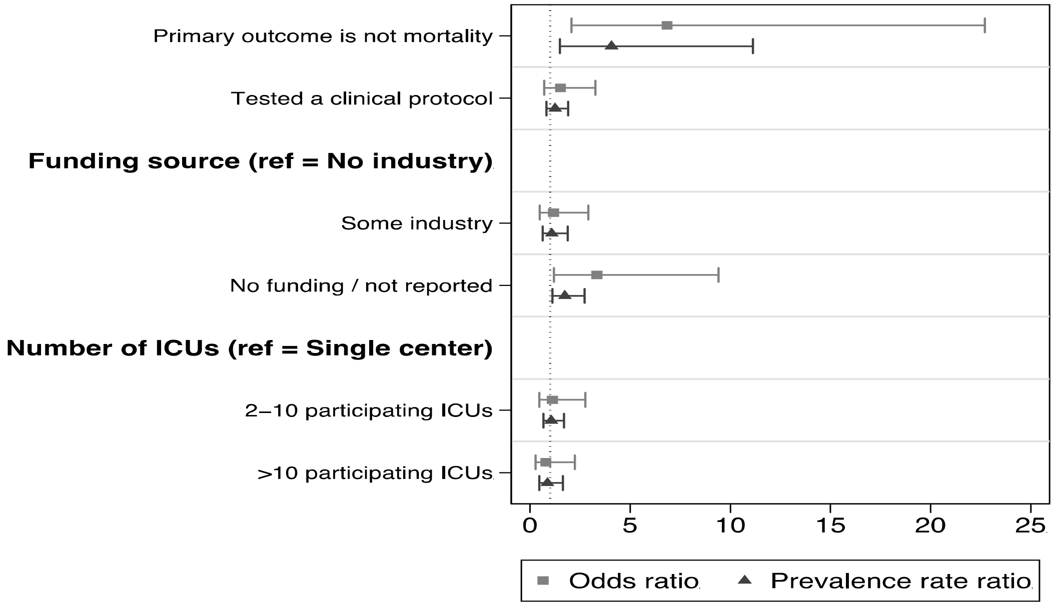
Table 2.3. Characteristics of adult randomized clinical trials in critical care

Characteristic	N (%)	N (%) with a positive primary outcome
Total	146 (100%)	54 (37%)
Funding		
No industry	80 (55%)	26 (33%)
Some industry	42 (29%)	13 (31%)
No funding / not reported	24 (16%)	15 (63%)
Single center	54 (37%)	25 (46%)
Multicenter	92 (63%)	29 (32%)
10 or fewer ICUs	40 (27%)	18 (45%)
11-25 ICUs	25 (17%)	6 (24%)
>25 ICUs	27 (18%)	5 (19%)
Type of intervention studied		
Protocol	71 (49%)	30 (42%)
Drug	59 (40%)	18 (31%)
Device/ monitoring	5 (3%)	1 (20%)
Other	11 (8%)	5 (45%)
Primary target patient populations		
General ICU	52 (36%)	30 (58%)
Sepsis spectrum	22 (15%)	0
Cardiac critical care	17 (12%)	7 (41%)
Acute lung injury/ ARDS	16 (11%)	2 (13%)
Unit of randomization		
Patient, surrogate, or family	137 (94%)	49 (36%)
ICU (cluster randomization)	9 (6%)	5 (56%)
Primary outcome (1 per trial, n=146 CCRCTs)		
Mortality (e.g., Hospital, ICU, 28-day)*	40 (27%)	4 (10%)
Infection related	33 (23%)	19 (58%)
Ventilation related	30 (21%)	13 (43%)
Quality (complications/adverse outcomes)	14 (10%)	7 (50%)
Organ failure	8 (5%)	1 (13%)
Composite outcome	7 (5%)	2 (29%)
Delirium	5 (3%)	2 (40%)
Hospital Discharge disposition (functional status)	3 (2%)	1 (33%)
Length of Stay	3 (2%)	2 (67%)
Smoking cessation	2 (1%)	2 (100%)
Quality of sleep	1 (1%)	1 (100%)
Most frequent secondary outcomes (multiple possible per RCT)		
Mortality		
ICU Mortality	47 (32%)	4 (9%)
In-hospital mortality	44 (30%)	2 (5%)
28 day	29 (20%)	4 (14%)
29-180 days	35 (24%)	5 (14%)
Ventilation		
Duration of MV	55 (38%)	12 (22%)
Ventilator free days	22 (15%)	6 (27%)
Length of Stay		
ICU length of stay	93 (64%)	12 (13%)
Hospital length of stay	71 (49%)	5 (7%)
Quality (Complications/adverse outcomes)	60 (41%)	14 (23%)
Infection related	36 (25%)	8 (22%)
Organ failure	17 (12%)	2 (12%)

The most common secondary outcomes across all RCTs were ICU (64%) and hospital (49%) length of stay (Table 2.3).

Of the 122 (84%) trials that disclosed the funding source, 34% reported receipt of industry funding, and 66% reported no industry funding. There was no relationship between industry funding and the probability that a trial would be positive (33% vs. 31%, p = 0.9). The remaining 24 trials did not disclose any sources of funding, and these were more likely to be positive (63%, p = 0.005 for comparison with all studies reporting funding sources). Single center RCTs (n=54) were less common than multi-center RCTs (n=92). However, multi-center RCTs were less likely to be positive and the rate decreased as the number of participating ICUs increased (p=0.03) in univariate analyses. In the multivariable regressions, RCTs that did not report any funding source (OR=3.3, 95% CI: 1.2-9.4) and RCTs that did not study a primary mortality outcome (OR=6.8, 95% CI: 2.1-22.7) were significantly more likely to be successful (Figure 2.2).

Figure 2.2. Adjusted associations of selected RCT characteristics with positive primary outcomes



Notes: RCTs that included measures of morbidity or other clinical measures in the primary outcome were not categorized as mortality trials.

Power or sample size were discussed in 135 RCTs (92%), however, only 68% of these studies cited prior research, a pilot study, or examination of other data (e.g., from the authors' center) to justify the inputs used in calculating the required sample size (Table 2.4). A CONSORT diagram portraying participant flow was reported in 119 RCTs (82%).

Table 2.4. Power and sample size characteristics of randomized clinical trials in critical care

Characteristic	Number of RCTs	n (%) positive
Total	*146*	
Included a consort diagram (patient flow)	119 (82%)	
Rationale for power parameters (e.g., baseline rate, predicted delta, expected time to event)	92 (63%)	
Type of outcome		
Binary outcome	101 (69%)	31 (31%)
Duration: (e.g., event free-days) or time-to-event	35 (24%)	16 (46%)
Rate (e.g., per 1,000 patient days)	7 (5%)	6 (86%)
Continuous	3 (2%)	1 (33%)
RCT stopped early	32 (22%)	
Futility	12 (8%)	
Safety	9 (6%)	
Recruitment / logistical issues	11 (8%)	
Power or sample size plan discussed, including cluster	*135 (92%)*	
RCT reported a targeted *a priori* sample size	130/135 (96%)	
Recruited < 95% of target or stopped early due to recruitment / logistical issues	20/130 (15%)	4/20 (20%)
Recruited 95-110% of target sample size or stopped early for futility	88/130 (68%)	36/88 (41%)
Recruited > 110% of target sample size	13/130 (10%)	4/13 (31%)
Stopped early for safety reasons	9/130 (7%)	

A total of 101 (69%) RCTs used a binary primary outcome. Of these, 40 examined a mortality outcome and 61 used other nonmortal outcomes (e.g., incidence of VAP). Twenty-three of the 40 RCTs with mortality as a primary outcome explained the rationale for their predicted treatment-associated mortality reduction**.** Thirty-four of these 40 RCTs reported the values for their power calculation and specified that they were superiority trials (i.e., powered for a specific treatment-associated mortality reduction). Of these 34 mortality endpoint superiority trials, three were positive (two only after pre-specified adjustment), and 11 (33%) had non-significant absolute treatment effects in the hypothesized direction that were larger than 3% (Figure 2.3**).**

Of the 61 RCTs with a primary nonmortal binary outcome, 47 were two-arm superiority trials and reported the predicted treatment-associated reduction they used for their power calculation (Figure 2.1**).** Of these 47 RCTs, 20 were positive and 27 were statistically non-

Figure 2.3. Expected versus actual treatment effect on mortality in 34 superiority trials where the primary outcome was mortality

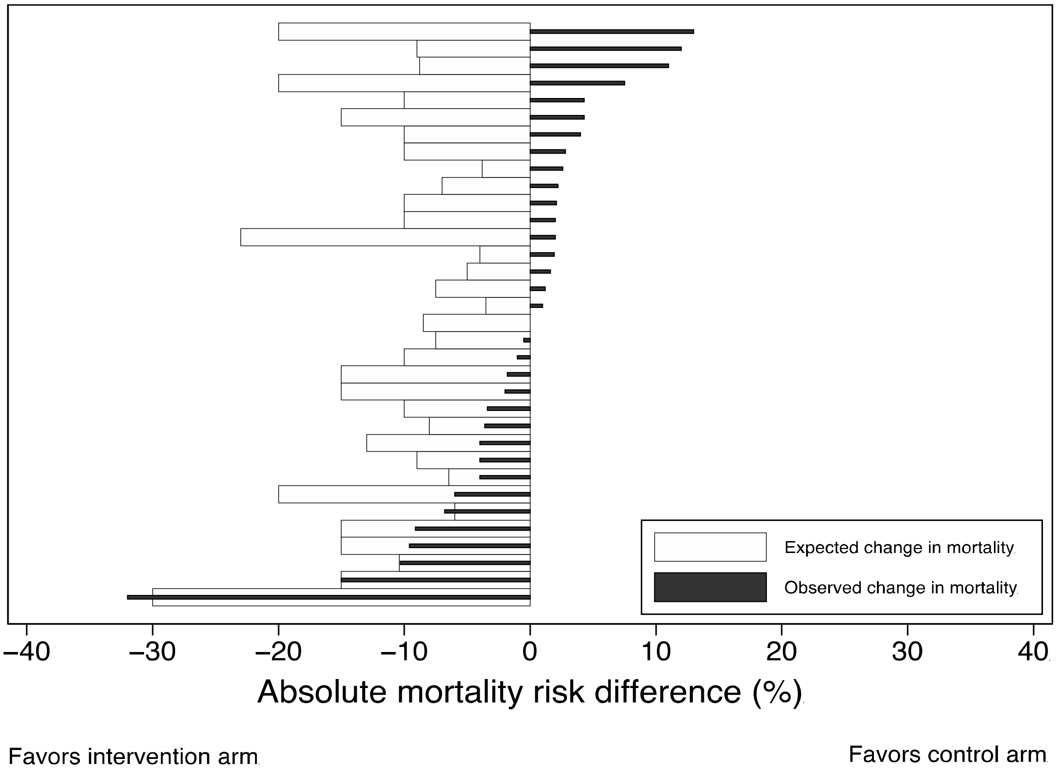
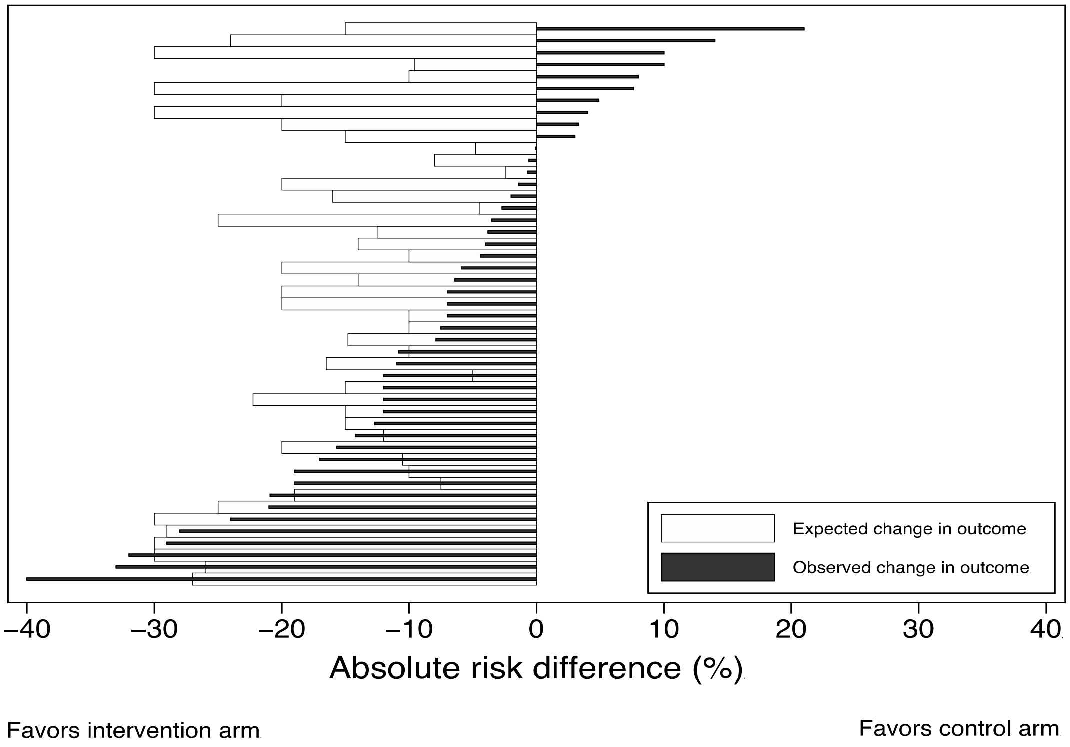


Figure 2.4. Expected versus actual treatment effect in 47 superiority trials with a binary non-mortal primary outcome



significant, of which 12 (44.4%) observed absolute treatment effects in the hypothesized direction that were larger than 3% (Figure 2.4).

Among the 33 superiority trials without adaptive control arms reporting expected control group mortality rates, the actual control group mortality differed from the expected value by 7.5% or more in 22 RCTs (Figure 2.5). Despite these frequent differences between expected and observed control group mortality rates, this rarely accounted for a study's inability to detect a given effect size as significant. For example, 12 (out of 30) negative mortality trials that tested for superiority could have detected a 10% mortality reduction with the observed control group mortality rate, compared with 13 such trials if the expected control group mortality had been observed (Figure 2.6). Among the 46 (of 47) nonmortal superiority trials with a binary endpoint in which expected control group rates were reported in the manuscripts, the actual control group rate differed from the expected value by 7.5% or more in 21 RCTs. Similar to the aforementioned

Figure 2.5. Expected and observed rate of mortality in control arms in RCTs that tested the effect of an intervention on mortality

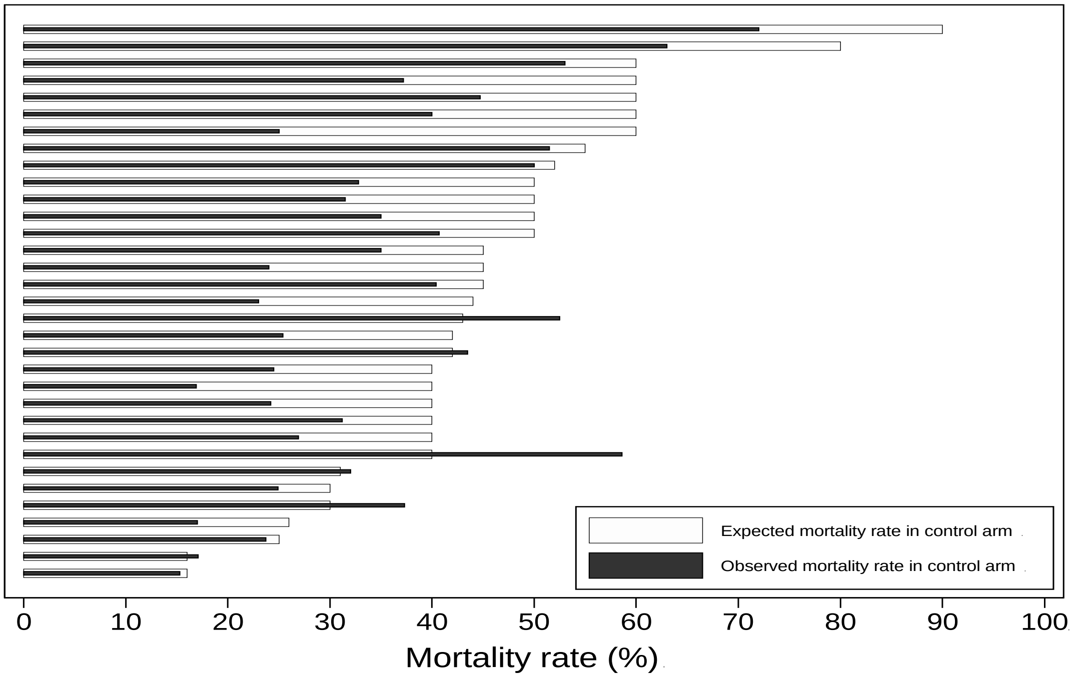
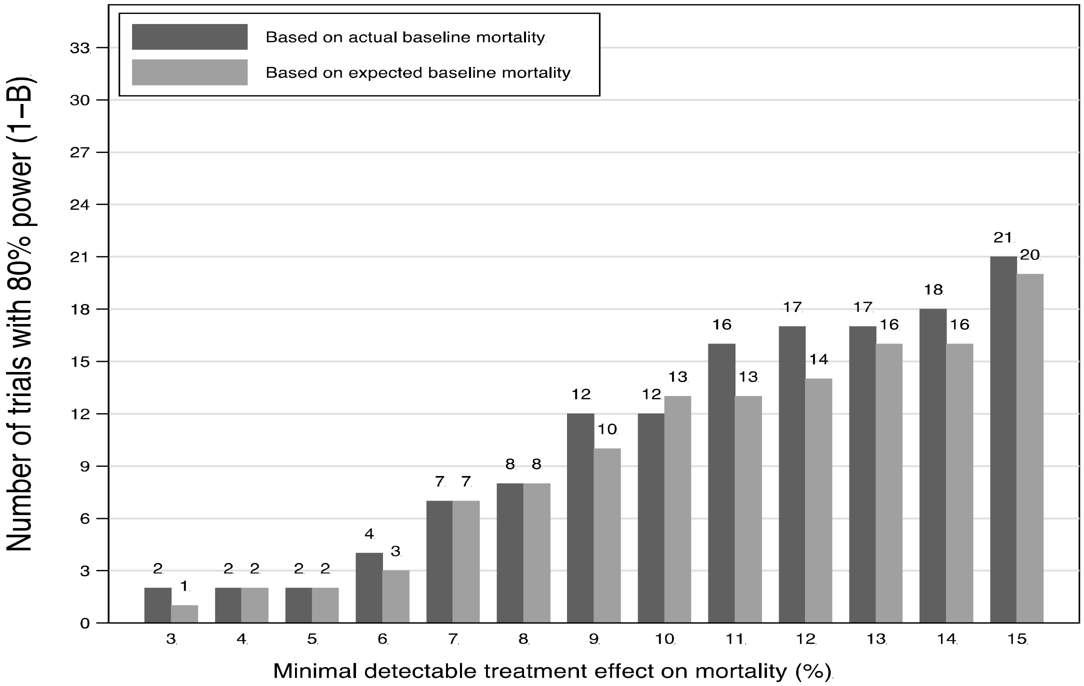


Figure 2.6. Simulation results of superiority trials where the primary outcome was mortality assuming 80% power to find a treatment-associated mortality reduction of 3 to 15%



results for mortality trials, misspecification of control group rates rarely accounted for a study's inability to detect a given effect size as significant.

Discussion

This contemporary study of 146 RCTs published in the leading medical and critical care journals yields several important findings. First, investigators choose a variety of primary outcomes for trials of ICU-based interventions. Some of this heterogeneity is appropriate given different anticipated effects of various interventions. However, the variation of endpoints selected even among trials using some form of a mortality primary endpoint suggests little agreement on the optimal outcomes in critical care. These data complement a prior study showing variability in ventilation-associated outcomes in critical care RCTs (Blackwood et al., 2014). This lack of standardized definitions and methods for assessing common outcomes poses challenges for comparing and understanding differences between RCTs, replicating results, and conducting meta-analyses.

Second, a majority of RCTs are “negative” in the sense that they do not demonstrate a benefit from the tested intervention. This is particularly true when mortality is the primary outcome (10% positive rate, or 5% if only crude rates are considered), with higher proportions of positive trials when other outcomes are used (13-100% positive rate) (Table 2.3). Of note, a 5-10% positive rate is roughly the rate that would be expected assuming a conventional type I error rate of 0.05. A prior review of RCTs in both adults and children published in the journal *Intensive Care Medicine* from 2000-2010 found an overall success rate of 48.8% (of 221 RCTs) (Latronico et al., 2013), somewhat higher than our observed rate of 37% (of 146 RCTs). Additionally, two reviews that focused on RCTs using mortality endpoints found success rates of 14% (10 of 72 RCTs published before August 2006) (Ospina-Tascon et al., 2008) and 18% (7 of 38 RCTs published from 1999-2009 in 5 major medical journals) (Aberegg et al., 2010), somewhat higher than our rate of 10%. Although it is possible that more trials are becoming negative over time, these differences may also be attributable to variability in the journals sampled and the eligibility criteria used to include RCTs. Because our study and all prior studies focused on published RCTs, the true rates of successful trials are likely even lower.

The high rate of negative trials does not, itself, suggest a problem; a majority of trials may "appropriately" fail to detect significant reductions in mortality. Such "true negatives" could arise if more interventions being tested are truly ineffective, as may occur when a discipline matures. Alternatively, such findings may be attributable to the fact that mortality in the ICU is heavily determined by physicians' decisions to withhold or withdraw life support (Garland & Connors, 2007), crowding out any plausible effect of an intervention. Finally, 10% or 20% of trials should be negative by chance alone even when power is set to 90% or 80%, respectively.

Nonetheless, the present study suggests that in many cases, critical care RCTs, and especially those studying mortal endpoints, have not been designed to identify realistic treatment effects. For example, we find that in a majority of negative RCTs, the results move in the predicted direction, often considerably so, yet fail to attain the predicted treatment effect upon which the study was powered (Figures 2.3 and 2.4). This provides contemporary evidence in support of the notion that investigators commonly select implausibly large treatment effects upon which to base sample size requirements (Aberegg et al., 2010). Although the problem of underpowered trials is certainly not unique to critical care, it does raise ethical concerns because such trials expose research participants to the risks and burdens of research without being (sufficiently) able to deliver on the purported benefits of expanding knowledge and improving future care (S. D. Halpern et al., 2002; Luce et al., 2004).

A third and related finding is that investigators commonly err in predicting the baseline event rate in their trials. With high-predicted background rates, large absolute risk reductions might seem plausible to investigators because they would reflect more modest relative risk reductions (Sackett et al., 1997). However, we find that control group mortality rates are often considerably lower than predicted, which could make such large effects improbable. For instance, it may be unreasonable to assume that an intervention predicted to bring mortality down to 30%, assuming a base rate of 40%, would also reduce mortality to 10% if the base rate turned out to be 20%. Thus, as the baseline mortality rate declines, there will invariably be diminishing marginal returns for any intervention – i.e., a lower proportion of potentially save-able patients.

Despite the possibility that over-predictions of control group event rates would contribute to critical care RCTs being negative, this appears to be only a minor piece of the problem. We found that even when large errors were made in predicted baseline mortality, this rarely changed whether a trial would or would not have detected a given difference as significant. This may be attributable to a counterbalancing phenomenon whereby as the baseline rate moves away from 50%, the sample size required to detect any given difference on an absolute scale decreases. Studies of secular declines in mortality rates for common pathologies, such as done with multicenter RCTs in sepsis (Stevenson et al., 2014) and acute lung injury (Spragg et al., 2010), could better inform control group mortality rates, and also guide selection of more reasonable treatment effects when designing future RCTs. Further, event-driven adaptive trial designs, such as utilized in the PROWESS-SHOCK trial (Ranieri et al., 2012), that adjust (by increasing sample size) to lower than expected mortality in the control group offer an attractive solution to this issue.

Additional strategies for improving trial success might include use of pre-specified covariate adjustment (Hernandez et al., 2004; Roozenbeek et al., 2010; Roozenbeek et al., 2009) [e.g., see Jansen and colleagues (Jansen et al., 2010)], larger target sample sizes, and more realistic and conservative treatment effect expectations (Scales & Rubenfeld, 2005) (Table 2.5). Additionally, innovative trial designs, such as Bayesian adaptive trials, may be particularly valuable for assessing drug therapies (Angus & van der Poll, 2013; Spragg et al., 2010). Regarding endpoints, some have questioned the conceptual propriety of using mortality as an endpoint for research or quality assessment on seriously or critically ill patients (Holloway & Quill, 2007). Although many experts believe that mortality is the ultimate patient-centered outcome for critically ill patients, others have called for greater use of nonmortal clinical endpoints (Ferguson et al., 2013; Spragg et al., 2010). Unfortunately, nonmortal endpoints face several threats to validity including, but not limited to, ascertainment bias (measurement error) and the limits of commonly used statistical methods for addressing the competing risks and informative dropout attributable to high ICU mortality rates. Indeed, our observation that RCTs of nonmortal endpoints were more likely to be positive may be an artifact of these measurement and analysis problems. Ongoing methodological work designed to offer new critical care outcome measures that incorporate mortality into the assessment of ICU length of stay or post-ICU quality of life may ultimately offer optimal approaches for quantifying the effects of interventions in the ICU.

This study has limitations. First, we only calculated power and detectable differences for trials using binary endpoints. We considered methods to assess effect sizes of trials employing continuous or time-to-event outcome such as ventilator-free days or time to extubation. However, potential effect size cutpoints (i.e., Cohen's *d*, Glass's Δ or Hedges's g), are all based on assumptions of normally distributed data. Because we found these assumptions unrealistic for most critical care outcomes, and the inputs difficult, if not impossible to back calculate from the published findings, we limited our power assessments to trials using binary outcomes. Second, our review was limited to adult critical care RCTs published in 16 selected journals. Third, since we relied on published data (and online supplements when available), changes in journal requirements over time may have contributed to certain reporting omissions (e.g., funding information or CONSORT diagrams). Fourth, important design issues such as allocation concealment, blinding or masking and ascertainment bias were not assessed. Finally, while we implemented an exhaustive search with oversight from a medical librarian, it is conceivable that our search strategy did not identify all eligible trials.

In summary, we believe greater dialogue is needed to determine the utility of nonmortal outcomes to patients, providers, and payers, and to identify elements of trial design and analysis that are associated with the significance of results (Naylor & Llewellyn-Thomas, 1994; Spragg et al., 2010). Rather than abandoning RCTs, the results suggest opportunities for designing critical care trials more efficiently. Actionable first steps might include consensus building among the critical care community (including journal editors) regarding a minimum core outcome set (Blackwood et al., 2014; Young et al., 2012), methodological work to improve strategies for measuring these outcomes, and closer scrutiny of submitted manuscripts to ensure an “honest” power calculation, which should in turn encourage more realistic trial design.

In the subsequent chapters we examine some of these considerations by focusing on nonmortal outcome measurement and analysis.

Table 2.5. Selected recommendations for critical care trial design

Domain	Hypothesis	Recommendations to potentially improve design
Study population	Treatment-effect heterogeneity might lead to a diluted effect estimate because while interventions work for certain patients, others are too sick and/or have too many competing risks for death for singular interventions to be of benefit.	• Stratified randomization. • Pre-specified severity of illness adjustment when estimating treatment effects (Hernandez et al., 2004). • Stratification of trial results based on severity of illness at baseline (Kent et al., 2008; Kent & Hayward, 2007). • Adaptive trial designs (e.g., using biomarkers to stratify patients into more homogeneous subgroups (Angus & van der Poll, 2013), event-driven adaptive trials (Ranieri et al., 2012), or starting trials with several arms and then adjusting sample sizes (Friede & Kieser, 2006) or narrowing arms based on observed interim safety and efficacy data (Lewis et al., 2013)).
Participant accrual and retention	RCTs are sufficiently powered but patient attrition leads to appreciable post-randomization losses so that the intention-to-treat analyses are highly conservative.	• Incorporation of patient attrition estimates when making sample size calculations. • Improved models of informed consent (Scales, 2013) and potentially incentives for research participation (S. D. Halpern, 2011a).
Statistical power calculations	Even when the target sample size is achieved and retained, RCTs may be insufficiently powered to detect relatively small but important effects on appropriate outcomes.	• Increased meta-studies to better inform control arm event rates (e.g., (Stevenson et al., 2014)). • Use of more realistic and conservative predicted treatment effects when estimating sample sizes. • Use of continuous outcomes when possible. • Reconstruction of binary endpoints into categorical endpoints to improve statistical efficiency (McHugh et al., 2010; Roozenbeek et al., 2010).
Outcome	Outcome measures are inappropriately specified or analyzed.	• Consensus development among trial groups and intensivists about follow-up periods and definitions of outcomes for specific conditions to support comparisons across trials (e.g., meta-analysis) (Blackwood et al., 2014; Young et al., 2012). • Novel methods for handling right-censoring due to deaths in analyses of quality of life and other nonmortal outcomes (Rosenbaum, 2006).

The following references contributed ideas presented in this table: (Aberegg et al., 2010; Angus et al., 2010; Annane, 2009; McAuley et al., 2010; Ospina-Tascon et al., 2008; Reade & Angus, 2009; Rubenfeld & Abraham, 2008; van Meurs et al., 2008).

CHAPTER 3. NONMORTAL TRIAL ENDPOINTS: EMPIRICAL FRAMEWORK FOR A CASE STUDY OF INTENSIVE CARE UNIT LENGTH OF STAY

Rationale for examining nonmortal endpoints

The objective of this chapter is to summarize the conceptual framework and empirical methods used for the statistical simulation studies that are reported in Chapters 4 and 5. These two chapters focus on nonmortal endpoints, with ICU LOS as a case illustration.

As a sizeable proportion of ICU-based RCT study subjects die before the trials are completed (Mebazaa et al., 2016), a mathematical tension has emerged for future trial design. If researchers continue to choose mortality as an endpoint (thus powering trials for a difference in proportions), trials will either have to enroll substantially larger patient populations, or pursue increasingly larger treatment effects. The former option potentially limits the feasibility of trials, and the latter option increases the risk of missing small, but clinically important treatment effects.

Accordingly, several thought leaders and trial consortiums in critical care have advocated the importance of validating patient-centered and clinically relevant nonmortal endpoints (Opal et al., 2014; Spragg et al., 2010; Young et al., 2012). However, lack of agreement concerning specific definitions and analytic methods for these outcomes may limit the external validity and applicability of RCT findings (Blackwood et al., 2014; Contentin et al., 2014). Therefore, Chapters 4 and 5 seek to contribute knowledge on the current scope of methods used to assess nonmortal endpoints, and illustrate potential modifications in interpretation that may achieve the stated goal of improving the design and interpretation of critical care trials.

Intensive care unit length of stay

In Chapter 2, we found that ICU LOS is the most frequently reported primary or secondary outcome among ICU-based trials. ICU LOS is a promising nonmortal trial endpoint for at least five reasons. First, LOS is easily measured from claims data and electronic health records. Second, LOS is important to patients and their families, whose quality of life is impacted by hospitalization and intensive care (Iwashyna, 2010). Third, LOS is relevant to all patients, in contrast to other common nonmortal ICU trial outcomes, such as ventilator-free days and organ-failure-free days. Fourth, LOS is a practical measure of resource allocation that can be quantified in economic terms (Cooke, 2012; Dasta et al., 2005; Kahn et al., 2008; Rapoport et al., 2003). Fifth, continuous outcomes such as LOS generate greater statistical power than dichotomous or categorical outcomes, thereby facilitating detection of effective treatments (Altman & Royston, 2006; McHugh et al., 2010).

Designing a simulation study of intensive care unit length of stay

In our conceptual framework of ICU LOS, we identified several processes, including overall mortality, mortality differences, and procedural factors that may affect the total duration of ICU stay of a study cohort. In Chapters 4 and 5, we isolate these processes and generate subsequent distributions of ICU LOS to examine how the interpretation of treatment effects for continuous, duration-based endpoints can be challenged by issues related to informative censoring from mortality and measurement error. While each analysis focuses on distinctly different aspects of interpreting LOS treatment effects, the empirical approach for each is interrelated. Therefore, we present the general summary used for each analysis here to avoid redundancy within each chapter.

Both of these research studies rely on the analysis of simulated (i.e., hypothetical or artificial) ICU LOS distributions. Simulating a LOS distribution that reflects potentially real-world settings is not a straightforward process. For example, the ICU LOS for a given trial population consists of a heterogeneous mix of subjects who died and those who survived. Consequently, ICU LOS may represent two very distinct clinical outcomes; for some subjects, ICU LOS may represent time to death, and for others, represents time to clinical improvement. Understanding how changes in mortality (Chapter 4) or measurement error among survivors (Chapter 5) can impact the interpretation of a LOS estimate requires that the analyst can isolate these processes and to keep other processes identical so that any remaining sources of variability are removed.

To achieve the inferential aims of each chapter, we cumulatively employ three different approaches to data generation. For each of the data generation processes, we use the *survsim* package in STATA (College Station, Texas), which uses a competing risk multistate model to generate time-to-events (Beyersmann et al., 2009; M. J. Crowther, 2011; M.J. Crowther & Lambert, 2012). In each setting, we manipulate cause-specific hazards for death and discharge to generate ICU LOS. Utilizing this data generation process, we then assume an intervention could modify the overall LOS distribution through three mechanisms based on the goal of the simulation:

1. LOS among survivors and LOS among decedents are independent, such that a change in one does not impact the other.
2. LOS among survivors and LOS among decedents are correlated, such that an effect of an intervention among survivors could reduce the fraction of mortality (i.e., mortality rate) in the treatment arm because patients are discharged faster, but an intervention does not impact the time-do-death (i.e., hazard) among those who die.
3. LOS among survivors and LOS among decedents are correlated, such that an effect of an intervention that modifies the time-to-death among decedents could also modify the time-to-discharge, as saved patients would shift into the risk-set for discharge and then be exposed to the discharge hazard.

Each approach has merits when the goal is to understand how different sources of bias can manifest in treatment effect estimates. We employ the third approach in Chapter 4. In this analysis we are precisely interested in how changes in mortality can modify summary estimates and statistical comparisons of ICU LOS. While the duration outcome in these chapters is LOS, such a decision is semantic, such that any duration could be selected for the hypothetical simulation (i.e., duration of mechanical ventilation). Accordingly, in this analysis we assume that an intervention does not change LOS among those patients who would always survive. We then modify the cause-specific hazard for death and quantify how statistical comparisons of LOS change using a range of statistical models found to be used in the systematic review. Therefore, we are able to gain insight into how mortality can modify LOS distributions and thus, statistical comparisons.

In Chapter 5, the goal is to estimate how factors extraneous to a subject's clinical state can also impact treatment effect estimates. We utilize the first and second data generation frameworks above to achieve this goal. We chose to use these two settings because the choice of either permits differences to exist between the total number of survivors, who are all potentially exposed to this form of measurement error. First, we assume that there is a group of patients who will always die regardless of their intervention arm, as well as a group who will always survive. A clinical analog might be an intervention that reduces the rate of an infection that is not life-threatening. As a result, a shorter LOS is observed among treated patients since the control arm, on average, requires additional ICU care until resolution of infection. However, survival is not affected. We call this the '*principal stratification*' framework as we only modify LOS among those in the cohort who are discharged alive from the ICU. In the second setting, we alter the setting above such that the intervention, in reducing ICU LOS, indirectly reduces overall mortality in the treatment arm among a subgroup of patients who would have died in the ICU in the previous framework. Thus, we conceptualize the trial population to be comprised of those who will always survive, those that would have died in the ICU but since they were discharged faster survive their ICU stay, and those who would always die during their ICU stay. The intervention only impacts the first two patient types, thereby reducing the overall ICU mortality rate, but having no impact on the time to death among those who die. Stated another way, such an intervention only passively reduces ICU mortality by reducing risks associated with being in the ICU (e.g., sepsis, blood stream infection) because treated subjects are discharged faster. This aligns with the conventional '*competing risk*' framework.

CHAPTER 4. HETEROGENEITY IN THE DEFINITION AND ANALYSIS OF INTENSIVE CARE UNIT LENGTH OF STAY IN CRITICAL CARE TRIALS

This chapter has been submitted for publication. The suggested citation is:

> Harhay MO, Ratcliffe SJ, Small DS, Suttner L, Crowther MJ, Halpern SD. Heterogeneity in the definition and analysis of intensive care unit length of stay in critical care trials. (Under review)

Introduction

This chapter focuses on empirical and interpretive challenges that may arise when interpreting ICU LOS when mortality is differential between study arms. Since critically ill patients commonly die, comparisons of nonmortal endpoints must properly account for these deaths either empirically or conceptually. Otherwise, the truncation of follow-up or censoring from death may cause nonrandom missing outcome data, potentially eroding the assurance of unbiased inference and thus interpretation when summarizing differences between study arms (Brock et al., 2011; Hernán & Robins, 2016; McConnell et al., 2008; Schoenfeld et al., 2002).

While some prior work has considered statistical models to account for this potential bias (Checkley et al., 2010; Chiba & VanderWeele, 2011; Deslandes & Chevret, 2010; Hayden et al., 2005; Resche-Rigon et al., 2006; Yang & Small, 2016), this paper focuses on understanding how nonmortal outcomes (using ICU LOS as an example) are examined in practice, and on gauging the potential impact of small, non-significant treatment-associated mortality effects on the results of typical analyses of nonmortal trial outcomes. To accomplish these goals, we extend the systematic review from Chapter 2 to assess the variability in the definitions and measurement of LOS in published RCTs, with specific attention to the methods used by researchers to manage competing events (i.e., death) when comparing nonmortal endpoints between study arms. Second, we use statistical simulations to assess the biases that may be generated from the most commonly used analytic methods found in the review. Finally, we use these findings to guide recommendations for reporting and analyzing LOS as an outcome in ICU-based RCTs, with extensions to other longitudinal, nonmortal outcomes.

Methods

Systematic review

For the present analysis, we extended the database detailed in Chapter 2 by two years, such that it now spans the period from 01/2007 through 06/2015. For each trial, the abstractors

(MOH and LS) identified whether LOS was the primary or secondary outcome, the definition provided by the authors, the statistical methodology used to compare LOS between treatment arms, and how the LOS distribution was reported (e.g., survivors only).

Illustration of interpretive bias through simulation

Building conceptually from recent critical care simulation studies and expert roundtables (Iwashyna et al., 2015; Mebazaa et al., 2016; Sjoding et al., 2015), we designed three simulation settings to illustrate how different mortality effects that may occur in reality could impact the interpretation of LOS. In setting 1, there was no treatment-associated mortality reduction (i.e., we simulated a perfectly null effect). In setting 2, the intervention imposed a constant effect over time (i.e., proportional hazards) such that the probability of a mortality reduction was equal for all patients. In setting 3, we isolated the treatment-associated mortality reduction to the simulated patients with a LOS in the upper tertile (i.e., time-dependent treatment effect) so as to reflect the possibility that the treatment might help only the sickest patients who tend to have longer LOS (Moitra et al., 2016; Zimmerman et al., 2006). Although we simulate beneficial mortality effects of treatment, identical results would manifest had we imposed the mortality reduction on the control arm. Thus, the results of this approach also apply to cases of harmful treatment effects.

As outlined in Chapter 3, to isolate the impact of these potential treatment-associated mortality effects on the interpretation of LOS comparisons, the simulated data are set such that the intervention would truly have no effect on LOS if all patients survived. Specifically, the parameter controlling the treatment effect was set to zero in the discharge sub-model, and set to give the imposed mortality reduction (2.5% or 5.0%) in the death sub-model with administrative censoring at 30 days. Thus, any observed LOS effect would be due to chance or the bias produced by the mortality effect. Such bias could arise if the treatment extended patients' LOS by saving them, or if it lengthened time-to-death among some patients who nonetheless die.

We express the primary outcome of interest of all simulations as the "interpretive error rate," defined as the percentage of simulations in each setting reporting a statistically significant difference in LOS between the intervention and comparator arm. The expected interpretive error rate due to chance is 5% using two-sided statistical tests with alpha=0.05.

To enhance the range of trials to which our results may apply, we adjusted four parameters in each of the three settings. First, the control-group 30-day mortality was set to 30% or 10%, representing relatively high and relatively low in-hospital mortality rates for modern RCTs. Second, we imposed an absolute mortality reduction of 2.5% or 5.0% in the intervention arms. We chose these effects because they would be clinically important, but most ICU-based RCTs would fail to detect them as statistically significant. Third, we examined short (median 3 days, interquartile range [IQR]=1.5-4.5) and long (10 days, IQR=5-17) LOS distributions, guided by ICU-based RCTs where LOS was the primary outcome (Ali et al., 2011; Amrein et al., 2014; Casaer et al., 2011; Kerlin et al., 2013). Finally, we simulated the total sample size as 250 or 1000 patients (125 or 500 patients per arm). This approach yields 2 x 2 x 2 x 2 = 16 simulations to be run in each of the three settings. However, we only applied a 2.5% absolute mortality reduction in settings in which the control-group mortality rate was 10% because a single ICU intervention would be unlikely to produce a 5% absolute reduction in mortality from 10%, and if it did, this effect would commonly reach statistical significance. Simulated data were generated as outlined in Chapter 3 (approach 3) using the *survsim* package in STATA, and 1,000 Monte Carlo replicates were used in each simulation setting.

We did not examine the statistical properties of ICU free-days or of changing the valuation of LOS to be the longest LOS or never discharged because the valuation of death as a specific LOS value has subjective elements that are beyond the scope of this work.

Results

We identified 193 eligible RCTs among ICU patients from 2007-2015. Of these, 150 RCTs (78%) reported on ICU LOS, with 132 of these trials explicitly reporting ICU LOS as an *a priori* outcome. In 6 trials LOS was specified as the primary outcome, and in 126 it was specified as a secondary outcome.

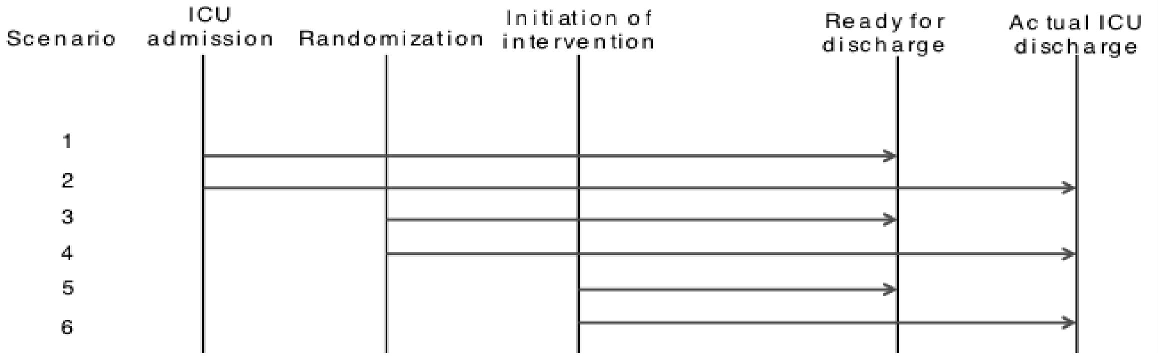
Definition and measurement of LOS

In 70 (47%) RCTs reporting on ICU LOS, insufficient details on the definition or measurement of LOS were provided to identify how ICU LOS was measured. Of the remaining 80 (53%) RCTs, at least either the start or end time was reported by the authors (n=54, 36%) or we believed one of these times could be reasonably deduced (n=26, 17%) based on the text and detail related to the study design or other trial outcomes. In 70 trials that had a reported or deducible “start time,” LOS measurement began at: a) the time of ICU admission (47%) b) the time of randomization or trial enrollment (34%), or c) the time of initiation of the intervention (7%). In the remaining 11% of trials, more than one start time was reported or two or more of these times appeared to overlap. In 70 trials that had a reported or deducible LOS “end time,” these times were specified as: a) ICU discharge and/or death (93%), or b) time of resolution of critical illness (7%). As a result, ICU LOS could represent six distinct durations based on the current literature (Figure 4.1). In addition, precision in the reported units of LOS was also quite variable, with 60 (40%) trials reporting LOS in 24-hour periods without rounding to the nearest day, and 77 (51%) trials reporting LOS as “days” without clarifying if days were calendar days or 24-hour periods. The remaining 9% reported LOS in hours (n=13).

Statistical analysis of ICU LOS

The statistical analysis used to compare LOS between study arms was unclear or not identifiable in 3 of the 150 trials and 10 trials reported to use ≥2 distinct statistical models. One

Figure 4.1. Variation in the reported definition of intensive care unit length of stay



trial assessed LOS as a binary outcome (i.e., prolonged LOS [LOS > 4 days] or not), and the remaining 146 trials treated LOS as a continuous outcome. Based on the information provided by the authors we concluded that 75 trials compared LOS using a non-parametric rank-based test, with 61 of such trials explicitly stating the use of this method. Similarly, we concluded that 51 trials used a linear parametric model; 32 of these explicitly stated this. Twenty-three trials used time-to-event methods, specifically a Cox proportional-hazards model (n=14) or a log-rank test (n=9).

Treatment of LOS among decedents

In the analysis of LOS, 92 trials (61%) reported the assessment of a pooled LOS distribution without discussion or statistical consideration of mortality, 12 (8%) studies assessed a stratified sample of survivors and 4 trials (3%) reported both a pooled and stratified result (these approaches are detailed in Table 4.1). The remaining studies reported at least one LOS value or approach that explicitly modeled or acknowledged mortality. A version of the ICU free-day outcome was reported in 19 (13%) trials, with the valuation of death (always equal to zero days) explicitly reported in ten trials. Nine trials (6%) changed the value of LOS to be the longest LOS (n=2) using a non-parametric model or to never be discharged in a time-to-event model (n=7). Nine trials (6%) (including two that also used an infinite time approach) reported using a time-to-event model explicitly stating to have censored LOS at the time of death while the remaining did not clearly report their censoring approach.

Table 4.1. Leading approaches used to account for mortality in the analysis of length of stay

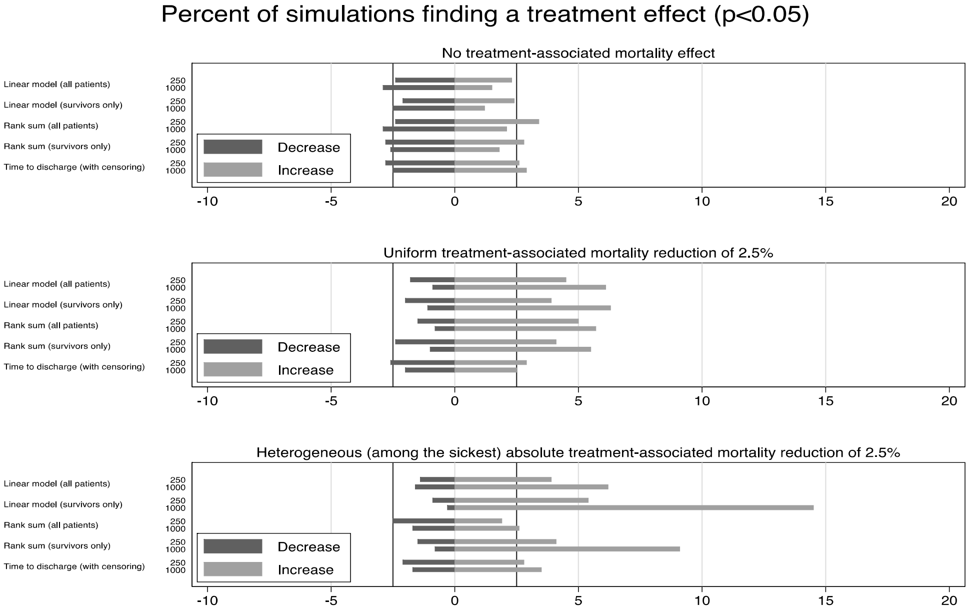
Approach	Conceptual and empirical issues	Consequence	Hypothetical trial scenario that would complicate approach
Contrast pooled LOS distribution of survivors and decadents together without acknowledging death.	LOS distribution is composed of treatment effects that may impact overall measures of dispersion but limited to a small number of patients.	Estimate from statistical test may be hard to interpret or misleading in isolation.	Patients saved by a treatment may experience a longer LOS that would impact the interpretation of the statistical comparison.
Construct and contrast a composite endpoint that includes both a value for death and LOS (i.e., ICU free-day metric where those who died are assumed to have 0 ICU free-days).	Valuing death inserts subjectivity to statistical analysis and fundamentally changes the causal question. ICU free-day metric does not have real world translation.	Estimates “net effect” of intervention. Disentangling effect of an intervention on either death or LOS is underpowered and limited by issues related to multiplicity.	Events of interest move in opposite directions, e.g., decreasing mortality and elongating LOS.
Contrast the LOS distribution among survivors only.	Survival may be affected by the treatment. Thus, it is a post-randomization variable. Conditioning on this can erode randomization inference and reduces study power.	Estimate from statistical test may be hard to interpret or misleading in isolation.	Saved patients are among the sickest, and thus experience longer LOS that would impact the interpretation of the statistical comparison.
Contrast time-to-live discharge in a time to event model and treat mortality as a form of non-administrative censoring.	Risk set subsequent the first death comprises a new subset of patients who have not previously died or been censored. Thus, balance of confounders assumed by randomization is potentially eroded. Statistical model assumes a “latent” LOS for censored patients – i.e., the LOS that patients who die are assumed to have had if they had lived, which is unknowable.	Estimate from statistical test may be hard to interpret or misleading in isolation. May cause selection bias unless assumptions of the model can be proven.	Heterogeneous treatment effects based on severity of illness or comorbidities.
Contrast time-to-live discharge in a time to event model and set the time to event to be infinite or longest possible LOS.	Patients do indeed have a chance of discharge in time-to-event model. Thus the statistical density of the time-to-event distribution is flawed, intentionally, and does not consider death as a competing event for discharge.	Estimate from statistical test may be hard to interpret or misleading in isolation.	Upon death, which will happen often in a trial, patients are removed from risk set.

Simulation study

The validity of our simulation approach was confirmed by the results in the control setting (setting 1) in which we assessed rates of interpretive errors using the most commonly reported methods of analyzing LOS data when there was no treatment effect on LOS or mortality. In this setting, LOS differences appeared nearly 5% of the time, as would be expected by chance (Panel A, Figures 4.2 & 4.3), and were not impacted by the overall mortality rate (not shown).

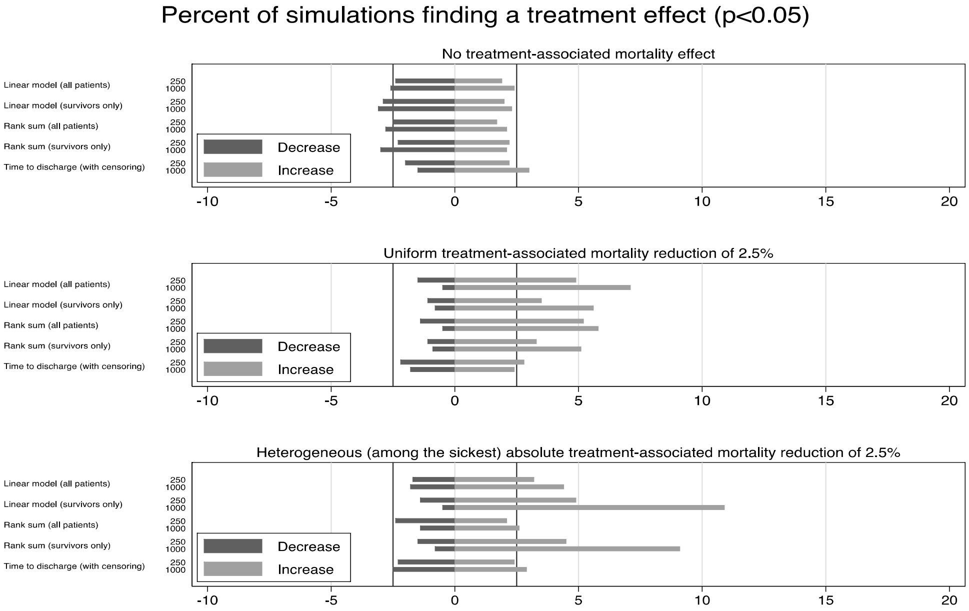
By contrast, all simulations were susceptible to interpretive errors when the treatment reduced mortality by 2.5% or 5.0%. The magnitude of bias depended on the total sample size, magnitude of the mortality effect, and the patients to whom the effect applied (i.e., uniform versus heterogeneous treatment effects) (Figures 4.2-4.6). When the mortality treatment effect was uniform (equal), we found that summary comparisons of the entire sample (which ignore differences between deaths and live discharges) performed the worst, with little difference between parametric and non-parametric comparisons. Similarly, we found high rates of interpretive errors across settings when these statistics were applied only to survivors. This effect was most pronounced when the treatment-associated mortality effect was isolated to the sickest patients (setting 3). Overall, time-to-event (discharge) analyses with censoring for death provided the lowest interpretative error rates. However, separate biases created by informative censoring may cloud the interpretation of results in time-to-event models (see discussion).

Figure 4.2. Percent of simulations exhibiting interpretive errors (primary setting, short LOS, treatment effect of 2.5%)



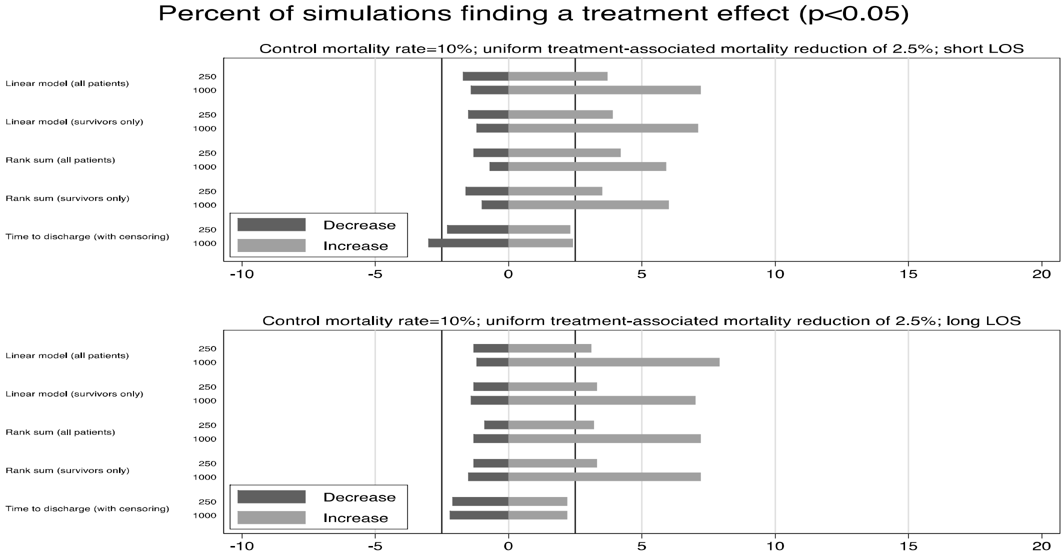
Notes: Based on chance approximately 5% of the simulations would be expected to be statistically significant, denoted by bars at -2.5% and 2.5% since we are using a two-sided test. Settings: Control group mortality rate of 30%, short LOS.

Figure 4.3. Percent of simulations exhibiting interpretive errors (primary setting, long LOS, treatment effect of 2.5%)



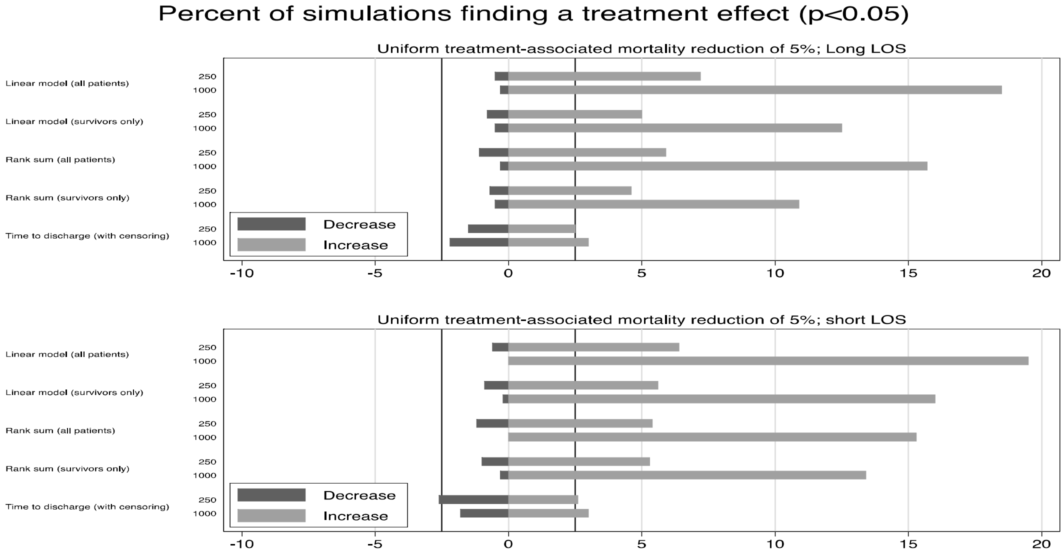
Notes: Based on chance approximately 5% of the simulations would be expected to be statistically significant, denoted by bars at -2.5% and 2.5% since we are using a two-sided test. Settings: Control group mortality rate of 30%, long LOS.

Figure 4.4. Percent of simulations exhibiting interpretive errors (sensitivity analysis, control group mortality of 10%, uniform treatment effect of 2.5%)



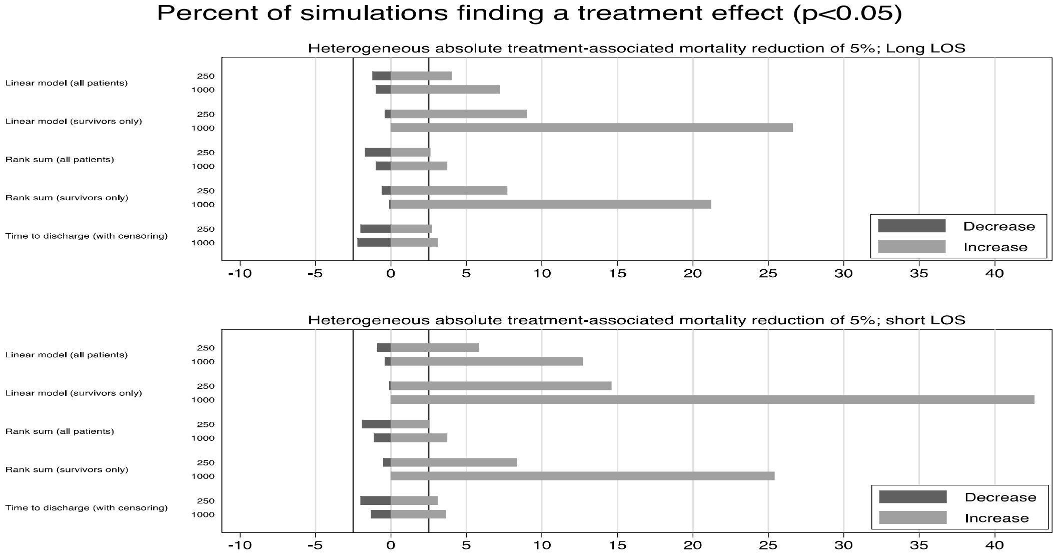
Notes: Based on chance approximately 5% of the simulations would be expected to be statistically significant, denoted by bars at -2.5% and 2.5% since we are using a two-sided test. Settings: Control group mortality rate of 10%, long & short LOS.

Figure 4.5. Percent of simulations exhibiting interpretive errors (sensitivity analysis, control group mortality of 30%, uniform treatment effect of 5%)



Notes: Based on chance approximately 5% of the simulations would be expected to be statistically significant, denoted by bars at -2.5% and 2.5% since we are using a two-sided test. Settings: Control group mortality rate of 30%, long & short LOS.

Figure 4.6. Percent of simulations exhibiting interpretive errors (sensitivity analysis, control group mortality of 30%, heterogeneous treatment effect of 5%)



Notes: Based on chance approximately 5% of the simulations would be expected to be statistically significant, denoted by bars at -2.5% and 2.5% since we are using a two-sided test. Settings: Control group mortality rate of 30%, long & short LOS,

Discussion

This study documents large variability in how LOS is defined and measured in critical care RCTs, and demonstrates the importance of accounting for the interplay between treatment effects on mortality and duration of illness when reporting nonmortal endpoints. Several specific results yield recommendations for the future measurement, analysis, and reporting of ICU LOS.

First**,** we identified a lack of consensus regarding how best to conceptualize ICU LOS. For instance, LOS was variably defined as time to discharge or death, as the time to live discharge, or in other ways. Similarly, few trials clarified the start or end time of their LOS measurement, and those that did showed a lack of consensus among three potential starting times (the time of admission, randomization, or intervention). Such variability limits the ability to compare interventions’ effects on LOS across trials. Similar problems arise from the noted variability across trials in the scale of LOS reporting (i.e., calendar days versus 24-hour periods without rounding versus hours), which may influence the magnitude of measurement errors.

Second, we identified at least five distinct analytic approaches that were commonly used to compare LOS between treatment arms in modern ICU-based trials (Table 4.1). These variations in statistical methods are not merely technical considerations, but rather lead to the testing of fundamentally different research questions. The most common approach is to contrast the overall LOS distributions in each treatment arm, without accounting for mortality. Our simulations suggest that this may generate misleading results in the context of small mortality effects that are commonly observed in trials, even when these effects are themselves not statistically significant. This degrades the ability to differentiate interventions that *seem* to lengthen LOS due to beneficial, albeit perhaps underpowered mortality effects, versus those that *truly* lengthen LOS without such corresponding benefits.

The other four commonly used approaches acknowledge the potential bias that differential mortality can create, but their use raises other interpretive challenges (Table 4.1). First, assessment of LOS among survivors reduces analytic sample sizes, which may be small in critical care trials to begin with. Additionally, such restriction can yield misleading results if a treatment shifts very sick patients from the “deceased” cohort to the “survived” cohort, where they may contribute an unusually long LOS (Lin et al., 2014). Setting 3 of the simulation showed that this approach can be especially problematic in the common cases in which treatment effects are not uniform (Iwashyna et al., 2015).

Second, investigators may use a time-to-event model that estimates the time to live discharge. Although censoring on death is likely superior to ignoring it altogether, such censoring assumes that death is random and non-informative. This assumption is almost certainly untenable (D. R. Cox et al., 1992), as patients’ acuities and comorbid conditions are related to both their probability of dying and their LOS if they survive. Thus, the probability of censoring may be time-dependent. If so, censoring could introduce bias despite randomization (Aalen et al., 2015). A third approach values death as a fixed LOS for decedents. The most frequent valuation approach uses an ICU free-day method, where LOS is set equal to the maximum follow-up time minus ICU LOS for live discharges and 0 for decedents (Schoenfeld et al., 2002). This approach quantifies the “net-benefit” of a composite clinical outcome, although it lacks a clear translation into clinical or economic benefit without separating the mortality and LOS components. Assessing potential interpretative errors when using ICU free-days would require a complex assessment of patient preferences as well as disparate treatment effects to fully evaluate their utility (e.g., increase in mortality among some patients and a decrease in LOS among others) and thus are not included in our current study.

A similar consideration regarding the “value” of death is raised in the fourth approach with changing LOS to be the longest observed LOS or treating the patient as never being discharged if they died during the study period. In a time-to-event framework this raises potential empirical issues due to the intentional distortion of the at-risk sample over time that needs to be better understood to avoid unintended introduction of another empirical bias. A more recently proposed approach that surmounts some of these conceptual problems is to code LOS as the longest possible LOS (infinite time), and use non-parametric tests to compare LOS distributions among treatment and control groups (Lin et al., 2014). Simulations using this approach suggest that it can accommodate a range of values for death, such as coding it at the 80th percentile of the LOS distribution or as the worst possible LOS (Lin et al., 2014). Thus, the approach enables investigators to assess the possibility that the conclusions to be drawn may be sensitive to how patients value death versus prolonged ICU stays. This approach is also flexible in that it enables investigators to estimate treatment effects on the median LOS, 75th quantile of the LOS, or any other point in the distribution (Lin et al., 2014). However, further experience with this approach is required to determine whether it will be acceptable to and understood by key stakeholders.

Until such experience is gained, and based on the systematic review and simulations presented in this manuscript, we recommend general standards for reporting LOS (Table 4.2) that are more general than previously published recommendations for free-day outcomes (Contentin et al., 2014) and broadly applicable to nonmortal outcomes. In addition, researchers may consider using more recently developed alternative statistical inference methods (Checkley et al., 2010; Deslandes & Chevret, 2010; Resche-Rigon et al., 2006), based on their inferential objectives in primary or secondary analyses. Finally, if longer follow-up after ICU discharge is available (e.g., 6-month or 1-year mortality), principal stratification methods may be considered to report nonmortal treatment effects (Chiba & VanderWeele, 2011; Hayden et al., 2005; Yang & Small, 2016).

Limitations

A limitation of the systematic review portion of this study is that the categorization of what authors had done in their trials was limited by differential reporting practices by authors as well as standards and requirements at different journals (e.g., publication of trial protocols) which may have impacted our ability to accurately document the LOS definitions and analytic methods. For example, it is possible that many trials utilized detailed and standardized definitions and measurements of LOS, but did not fully report them. This may be particularly true when LOS was a secondary outcome measure. While this is unlikely to change the overall interpretation of the results, this reality is important as the CONSORT standards for outcome reporting should apply equally regardless of the journal, and researchers pursuing meta-analyses and systematic reviews would encounter similar barriers when aggregating trial results. Thus, this limitation as well as the issues we highlight remain important to promote the generalizability and aggregation of knowledge from individual trials.

Second, though we sought to be exhaustive, it is possible that our search did not identify some published trials. Such omissions are unlikely to have been systematic, and so would not be expected to alter any of our conclusions.

Third, the simulation studies we presented were intentionally non-exhaustive of all potential ICU trial settings. Statistical simulation studies provide a unique lens to model hypothetical trial scenarios, and we have chosen a limited set of illustrative scenarios to understand if missing or truncated outcome data can bias how we assess interventions in the ICU. Many other scenarios are conceivable in an actual trial, such as those that include a treatment effect on LOS directly. These cases were not modeled in our simulations. However, the goal of this manuscript was to highlight prevalent problems that could be created due to small changes in mortality due to an intervention, rather than to provide an exhaustive accounting of the magnitudes of these problems in all possible scenarios. Future work may help to understand better the specific cases in which the magnitudes of the general biases we report are likely to be most extreme, as well as understanding how simultaneous changes in the risk of death, disease progression, and time-to-discharge among different patients can obscure or exaggerate effect sizes reported in trials. Finally, different data generation processes and assumptions may lead to different interpretations of the simulations.

Conclusion

Although ICU LOS is commonly used as an outcome measure in ICU-based RCTs, it is inconsistently reported and analyzed. Problems with heterogeneity of outcome use and definition are not limited to critical care, as documented in prior assessments of Cochrane reviews and ClinicalTrials.gov entries (Hirsch et al., 2013; Tovey, 2011). The present study shows how these choices may impact the interpretation of trial results. While challenging empirically and conceptually, we propose that researchers could employ some simple practices in reporting trial results to aid in their interpretation and synthesis. More granular reporting of mortality throughout the duration of follow-up, with reporting at the exact time the nonmortal measure is assessed, would help assess the risk of biased interpretation. Employing predefined secondary analyses with novel statistical approaches, such as the aforementioned rank-based method (Lin et al., 2014) and joint modeling approaches (Deslandes & Chevret, 2010), would enable experience to be gained with these methods so as to determine whether they ought to become standard. Finally, even very basic, but often ignored practices such as reporting the start- and end-time used to define LOS and the values of LOS applied to those patients that die during follow-up would greatly improve the interpretation of nonmortal endpoints in ICU-based RCTs.

Table 4.2. Recommendations for reporting and analyzing nonmortal endpoints in critical care trials

Domain	Problem identified in review	Recommendation
Measurement	Trials reporting start and end times varied in their definitions, and many do not report any definition at all.	Clear indication of the start and end time of the LOS measurement.
	Trials predominantly report LOS in "days." Calendar days and 24-hour periods are different, and can further vary based on the abovementioned issue of start and end times. This potentially adds measurement error.	Granularity and specificity in the measurement of LOS.
	A version of the ventilator free-day was used for ICU free-days, however, the treatment of death and follow-up period was not well defined in certain trials.	Detailed definition of composite outcomes that include LOS.
Analysis	Many trials simply state that nonparametric or parametric statistical models were used without any further detail. It is unclear in some trials which model was used to generate the p-value.	Stating exact model used to compare LOS between study arms.
	Similar to the issues noted for ICU free-days, values applied to death are sometimes used but not clearly stated.	Clearly stating assumptions of statistical analysis related to the treatment of LOS among decedents (e.g., censoring).
	The analytic sample assessed is not always the full trial population. If survivors only are analyzed it may not be clear which mortality cut-off was used to define this group (e.g., ICU, hospital, or 28-day mortality).	Cleary stating which patients were included in the analysis.
	Mortality is often reported at a few discrete time points (e.g., 28 or 60 days) or without clarity of total follow-up time (e.g., ICU mortality). This makes it difficult or impossible to assess trials for the potential of interpretive bias in the reporting of nonmortal endpoints if non-differential mortality occurs.	Report mortality rates at more granular time periods (e.g., 7, 14, 21, and 28 days).
	Most trials do not execute sensitivity analyses using advanced statistical methods.	Though the ideal or “correct” method for statistical inference is unclear, utilizing secondary methods such as competing risk, principal stratification, or joint statistical models can help researches assess their results.

CHAPTER 5. MEASUREMENT ERROR IN INTENSIVE CARE UNIT LENGTH OF STAY ESTIMATES DUE TO PATIENT FLOW

This chapter has been submitted for publication. The suggested citation is:

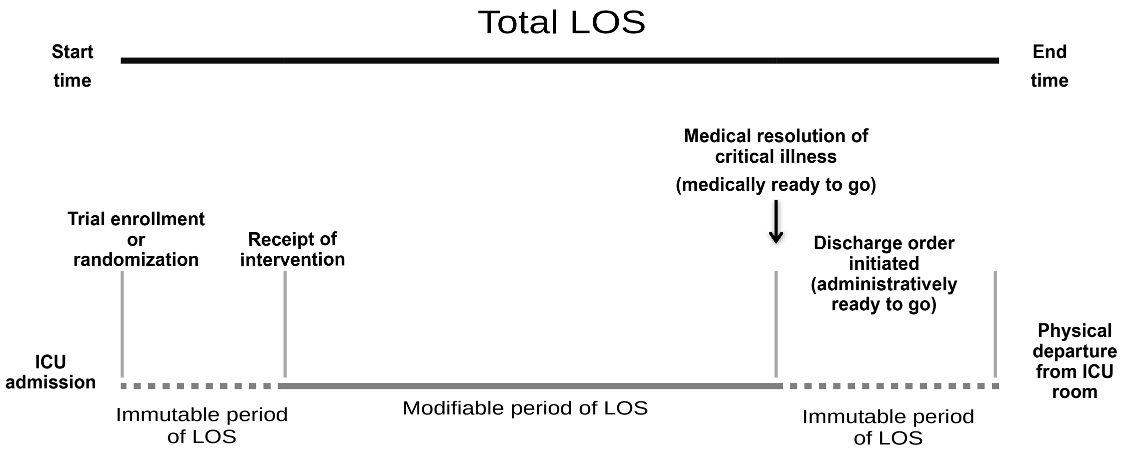
> Harhay MO, Ratcliffe SJ, Halpern SD. Measurement error in intensive care unit length of stay estimates due to patient flow. (Under review)

Introduction

In Chapter 4 we focused on potential bias and misinterpretation due to differential mortality. In this chapter we focus on another issue. Specifically, when evaluating the effect of a clinical intervention on LOS, one wishes to determine the intervention's impact on the time required for patients to become clinically ready for ICU discharge. Yet ICU-based RCTs typically report total LOS, defined as time until actual discharge. Because most ICU discharges entail patients transitioning to a step-down unit or general ward, factors such as bed availability and clinical rounding schedules may impact actual discharge time, independent of patients' illnesses or the interventions they receive (Wagner et al., 2013).

In this paper we examine the epidemiology and implications of such '*immutable time,*' which is termed as such because this time cannot plausibly be affected by most ICU-based interventions such as pharmacotherapies or ventilation strategies (Figure 5.1). We first perform a systematic review of ICU-based RCTs to identify studies that have defined ICU LOS using discharge readiness as an end time rather than actual discharge. Next, using statistical simulation informed by our own RCT and administrative data, we quantify the extent to which immutable time may actually bias estimates of treatment effects across a range of possible trial scenarios.

Figure 5.1. Decomposition of length of stay in an intensive care unit


Total LOS
Start time
End time
Medical resolution of critical illness (medically ready to go)
Trial enrollment or randomization
Receipt of intervention
Discharge order initiated (administratively ready to go)
Physical departure from ICU room
ICU admission
Immutable period of LOS
Modifiable period of LOS
Immutable period of LOS

Methods

Empirical framework

Information or measurement bias is a type of bias that results from measurement error (Lash et al., 2009). We define immutable time as a research bias because the resulting measure of LOS erodes the precision of statistical comparisons between study groups with the potential to obscure small, but clinically relevant treatment effects, or to suggest a treatment effect that does not exist. Below, we provide a framework for this problem. Specifically, we assume that immutable time is a non-differential measurement bias of a continuous (but non-normally distributed) variable.

The statistical model, $Y_i = f(X_i) + \epsilon_i$, represents a common analysis conducted in RCTs to quantify the effect of an intervention, where Y is ICU LOS, and X is a binary indicator for treatment arm generally measured without error, and ϵ_i indicates stochastic error. A parameter, β, is used to quantify the difference between the intervention and control LOS. For example, in the linear model $f(X_i) = \alpha + \beta X_i$. However, in lieu of LOS based on patient readiness, Y_i, the value of ICU LOS with immutable time, $\widetilde{Y_i}$, is measured, such that, $\widetilde{Y_i} = Y_i + \nu_i$, where ν_i is the immutable time unaffected by treatment arm. We assume ν_i is independent of X_i and ϵ_i. Under this framework, the additional bias term, ν_i, results in the following, $\widetilde{Y_i} = Y_i + \nu_i = f(X_i) + (\epsilon_i + \nu_i) = f(X_i) + \theta_i$, where θ_i is an error term that is biased by measurement error. In some trial designs early immutable time may also occur between the time of ICU admission and exposure to an intervention (Figure 5.1), and would effectively increase the mean and variability of the immutable time distribution.

Given that immutable time is essentially random and likely independent of treatment arm assignment, the analysis $Y_i = \alpha + \beta X + \epsilon_i$ (i.e., a linear model) should produce a correct estimate of β quantifying the difference (treatment effect) between treatment with a normally distributed outcome and normally distributed measurement error. The statistical explanation of why non-differential measurement error of an outcome (Y_i) would not bias the estimated treatment effect point estimate ($\hat{\beta}$) in linear model is as follows:

$$
\begin{aligned}
\hat{\beta} &= \frac{Cov\left(\widetilde{Y_\imath},\ X_i\right)}{Var(X_i)} \\
&= \frac{Cov(Y_i + v_i, X_i)}{Var(X_i)} \\
&= \frac{Cov(\alpha +\ \beta X_i + \epsilon_i + v_i,\ X_i)}{Var(X_i)} \\
&= \frac{Cov(\tilde{\alpha}, X_i)}{Var(X_i)} +\ \beta \frac{Cov(X_i, X_i)}{Var(X_i)} +\ \ \beta \frac{Cov(\epsilon_i, X_i)}{Var(X_i)} +\ \beta \frac{\square ov(v_i, X_i)}{Var(X_i)} \\
&=\ \beta \frac{Var(X_i)}{Var(X_i)} \\
&= \beta
\end{aligned}
$$

However, since neither ICU LOS nor immutable time, based on our empirical distributions (Figures 5.2-5.5) is normally distributed, it is unclear if this attenuation of the treatment effect will persist as variance may change with less predictability. Specifically, since classical measurement error has focused largely on linear regression models with normally distributed measurement error, it is not clear how immutable time could bias LOS treatment effect estimates.

With small treatment effects, it is possible that the measurement error will obscure treatment effects due to the resultant reduction in power. In other words, the variance will be larger than was presumed in calculating the required sample size. Further, as the size of a treatment effect increases, the more favorable signal-to-noise ratio would tend to mitigate the impact of this extra variance. Thus, we also hypothesize that the reduction in power due to immutable time would be most important in studies in which clinically relevant treatment effects are often numerically small, as is commonly true in ICU-based RCTs (Aberegg et al., 2010; Harhay et al., 2014; Rubenfeld, 2015).

Data analysis

Literature review

To identify trials accounting for time-to-discharge readiness in practice, we utilized the same data abstracted for Chapter 4, focusing specifically on those reporting the end time of their ICU LOS measurement.

Secondary data analysis

We examined daily, weekly, and yearly variation in discharge immutable time, defined as $LOS_{discharge}$ - $LOS_{ready-to-go}$ (indicated by a bed request for the patient on a general ward), at our institution using two data samples. First, we reassessed data from the SUNSET-ICU trial, an RCT comparing outcomes among patients whose nighttime management was overseen by senior intensivists who were physically present in the ICU versus at home and available by phone (Kerlin et al., 2013). Subjects included all patients admitted to the medical ICU (MICU) of the Hospital of the University of Pennsylvania during a one-year period (09/12/2011 to 09/12/2012) (Kerlin et al., 2013). In this trial, for patients readmitted to the MICU within the same hospitalization, only the first MICU admission was included. Next, we extracted data on all patients admitted to the same MICU from 2010-2012 to examine LOS variations over several years, and to assess all discharges, not just index admissions.

Simulation study

We performed a simulation study based on the results of our analyses of the SUNSET trial data and of administrative data sets. First, we generated a $LOS_{ready-to-go}$ distribution to approximate the ICU LOS distribution in the SUNSET trial using a Weibull distribution (median LOS of 2.5 days and IQR of 1.2-4.5). We assumed that the probability of death in the ICU was 10%, 20% or 40%.

As detailed in Chapter 3, the goal of a simulation study is to isolate a specific process, and to keep other sources of variability identical. To isolate the effect of immutable time in our simulations, only the time to discharge was manipulated. To capture the potential variability of different critical care interventions, we conducted a simulation study that generated data using both a principal stratification framework where the fraction of deaths was identical in study arms and a competing risks framework where the faster time to discharge resulted in a lower ICU mortality rate among the treated (without modifying the hazard for death) (see Chapter 3 for additional details). The objective of the two approaches was to assess how different LOS models, which treat or value LOS among deaths differently, may be affected by immutable time. For instance, in a time-to-event model LOS among those who die may be censored.

Second, we imposed hypothetical treatment effects of 0, 0.5, and 1 day at the median $\text{LOS}_{\text{ready-to-go}}$ for the treatment arm. Next, three immutable time distributions of increasing size (medians of 8, 16 and 28 hours in settings 1, 2, and 3, respectively) and generated from a gamma distribution were randomly added onto the $\text{LOS}_{\text{ready-to-go}}$ of survivors who were discharged (Figure 5.2). Thus, for each simulated trial there was an unbiased time-to-discharge-readiness (Y_i), and three LOS distributions with immutable time $(\widetilde{Y}_\iota)$. Setting 1 was based on late immutable time observed in the SUNSET Trial. Setting 2 was based on the longer distributions of late immutable time from our administrative data. Setting 3 assumed the combination of setting 2 and a median 12-hours of early immutable time between ICU admission and exposure to an intervention. We examined the impact of immutable time using four statistical approaches for comparing LOS that are commonly encountered in the literature as identified in Chapter 4. Specifically, the following methods identified in Chapter 4 were examined:

1) nonparametric comparison (i.e., comparing the entire distribution with a Wilcoxon Rank-sum test;
2) parametric comparison (i.e., comparing the means and standard deviation using ordinary least squares linear regression or t-test);

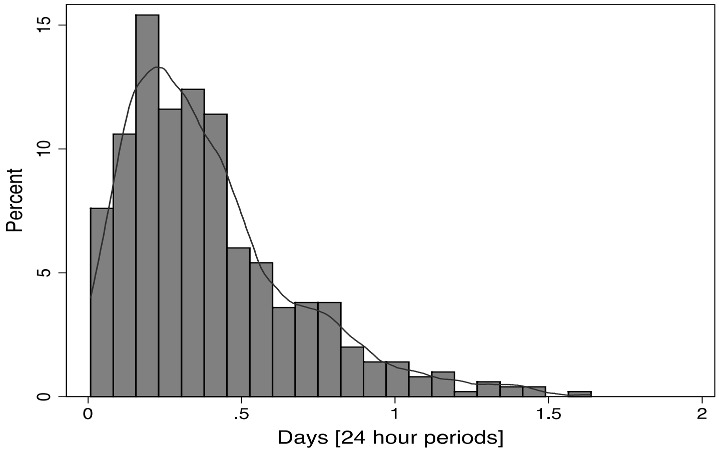
3) time-to-event model (i.e., Cox proportional hazards model of the time-to-discharge with mortality as a censoring event); and,

4) ICU free-days, where LOS is re-calculated as 30 days - LOS for survivors, and patients who die are given a fixed LOS of 0 free-days (Schoenfeld et al., 2002). ICU free-days are then compared using a nonparametric statistical model.

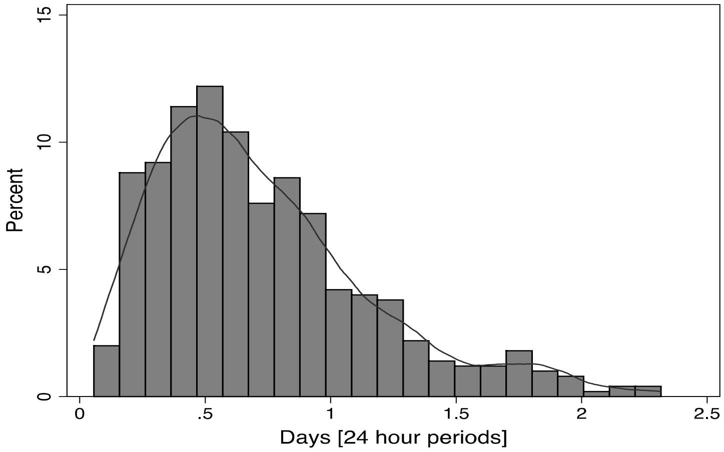
Total sample size was varied across settings (200, 400, 600, 800, 1000 and 1500 patients) to reflect the range of sample sizes observed in the majority of ICU-based trials identified in Chapter 2. All settings assumed an equal number of patients in each of the two study arms (1:1 randomization), and used one thousand Monte Carlo replicates. Administrative censoring occurred at 30 days. To quantify the potential effect of immutable time on the interpretation of analytic results, we summarized the percentage of times a two-sided statistical test in the three immutable time settings differed from the error-free-LOS$_{\text{ready-to-go}}$ at the α=0.05 level. Replicates were classified as false-positive if the intervention generated an effect on LOS at p<0.05 in the presence of immutable time but the same test yielded p≥0.05 in the absence of immutable time. Replicates were classified as false-negative in the reverse situation.

Figure 5.2. Sample immutable time distributions used for the simulation study

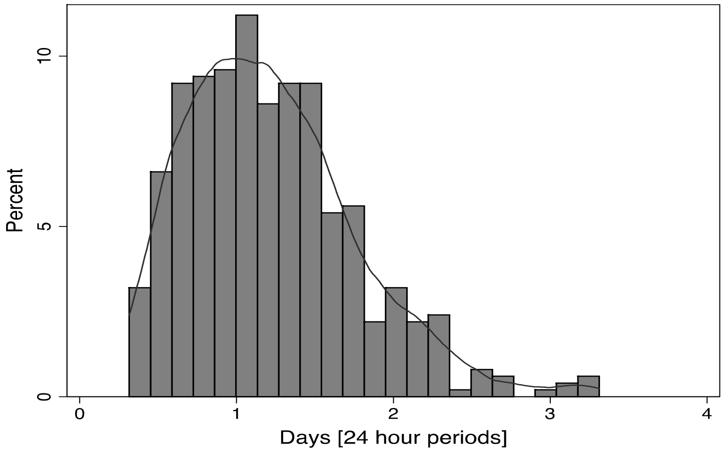
A) Setting 1



B) Setting 2



C) Setting 3



Notes: The figures are generated from a random data generation with a sample size of 500 (n=250 per treatment arm). A kernel plot is used to overlay the distribution. Data is generated using a Gamma distribution with the following settings:

A) shape=2 scale=0.205. Approximate median of 8 hours.
B) shape=3 scale=0.25. Approximate median of 16 hours.
C) shape=5.24 scale=0.235. Approximate median of 28 hours.

Results

Literature review

Trial investigators explicitly mentioned the use of a time-to-discharge readiness definition of LOS in 5 of 150 ICU-based RCTs (3%) in which LOS was a primary or secondary outcome (Table 5.1). Two trials used a ready-to-go time that was defined by clinical criteria, and two did not specify their criterion. In the fifth trial, the SUNSET trial, ready-to-go time was an administrative time-point indicated in the electronic medical record at the time the order for transfer from the ICU (indicating clinical readiness for discharge) was placed.

Table 5.1. Documented use of time-to-discharge readiness to compare ICU LOS between study arms

Author	Definition	Outcome
Casaer et al. (Casaer et al., 2011)	The duration of time in the ICU was defined as the time from admission of patients until they were ready for discharge. Patients were considered ready for discharge as soon as all clinical conditions for ICU discharge were fulfilled (i.e., no more need for vital-organ support and receipt of at least two thirds of caloric requirements as oral feedings) even if they were not actually discharged that day. The 'ready for discharge' day coincided with the actual day of discharge for all patients except for 104/2328 (4.5%) patients in the late-initiation group and 95/2312 (4.1%) patients in the early-initiation group.	Primary
Jakob et al. (Jakob et al., 2012)*	Length of study ICU stay was defined as time from randomization to being medically fit for discharge or transfer from the study ICU.	Secondary in both trials.
Tritapepe et al. (Tritapepe et al., 2009)	ICU length of stay (time meeting fit-for-discharge criteria). Patients were eligible for transfer out of the ICU when the following criteria were met: $SpO_2 \geq 90\%$ at an $FIO_2 \leq 0.5$ by facemask, adequate cardiac stability with no hemodynamically significant arrhythmia, chest tube drainage <50 ml h^{-1}, urine output ≥ 0.5 ml kg^{-1} h^{-1}, no i.v. inotropic or vasopressor therapy, and no seizure activity.	Primary
Kerlin et al. (Kerlin et al., 2013)	Length of stay as the time to request for a bed to a general ward.	ICU LOS (time-to-discharge) was primary, ready-to-go was a secondary definition.

*Two trials are reported in this publication

Secondary data analysis

Among the 1,149 MICU discharges to another hospital unit or ward among SUNSET trial participants during their index ICU admission, we observed a median ready-to-go time of 40.1 hours compared to 46.8 hours of time from admission to actual discharge (Figure 5.3). The median difference between ready-to-go time and actual discharge (i.e., immutable time) was 5.1 hours (IQR 2.7 to 8.9 hours). The 90th, 95th and 99th percentile differences were 14.2, 21.7 and 50.2 hours, respectively.

From administrative data of all MICU patients at our center in calendar years 2010-12, we identified 3,851 discharges from the MICU. The overall median immutable time was 7.0 hours (IQR 4.3 to 11.1, 90th percentile=21.6, 95th percentile=29.5), and this displayed considerable weekly, monthly, and yearly variation (Figures 5.4 & 4.5). More than half of all discharge requests were placed between 8 and 9 am, during which time the MICU's morning bed management rounds occur (Figure 5.6).

Figure 5.3. ICU length of stay ending at time of bed request and actual discharge among patients discharged in the SUNSET RCT

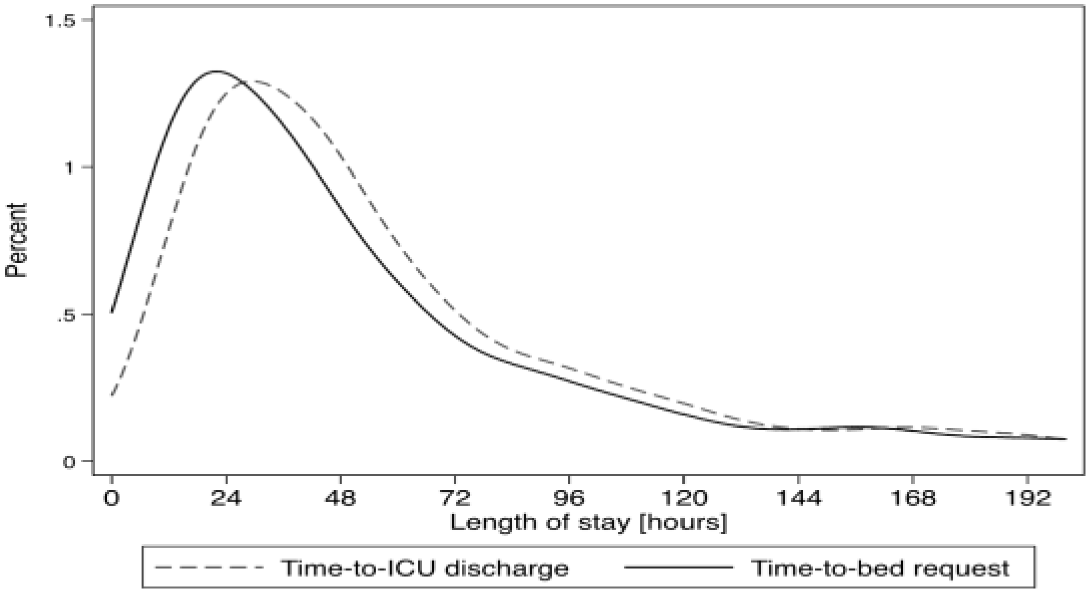
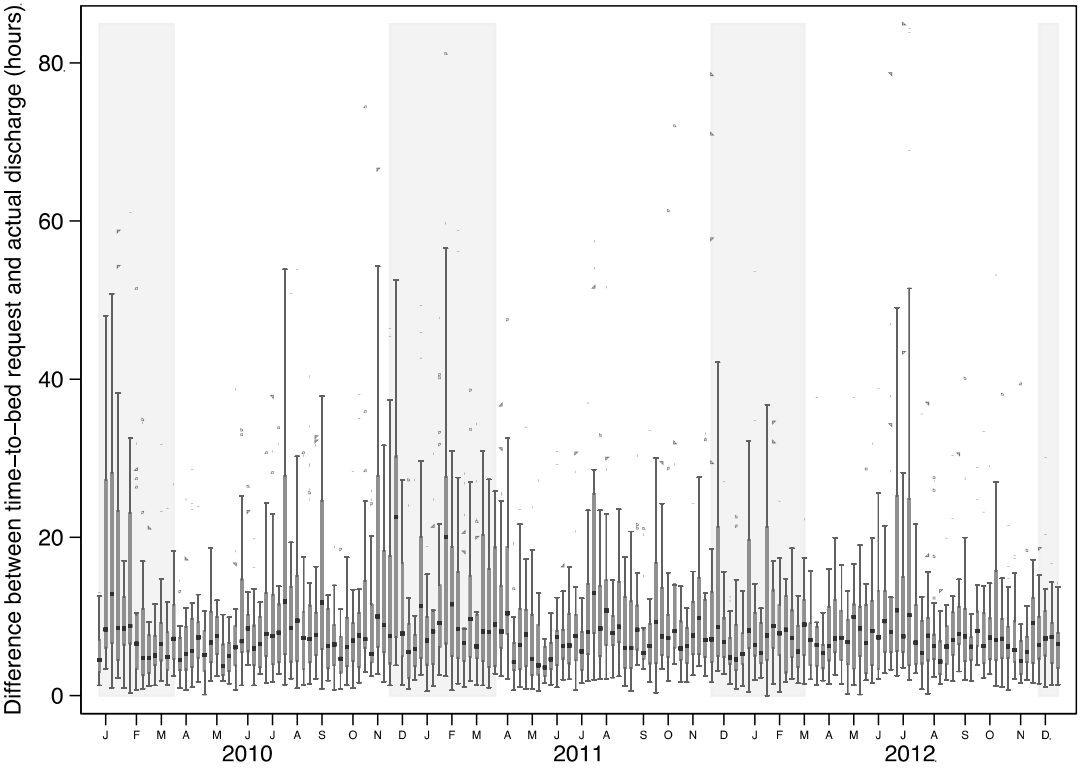
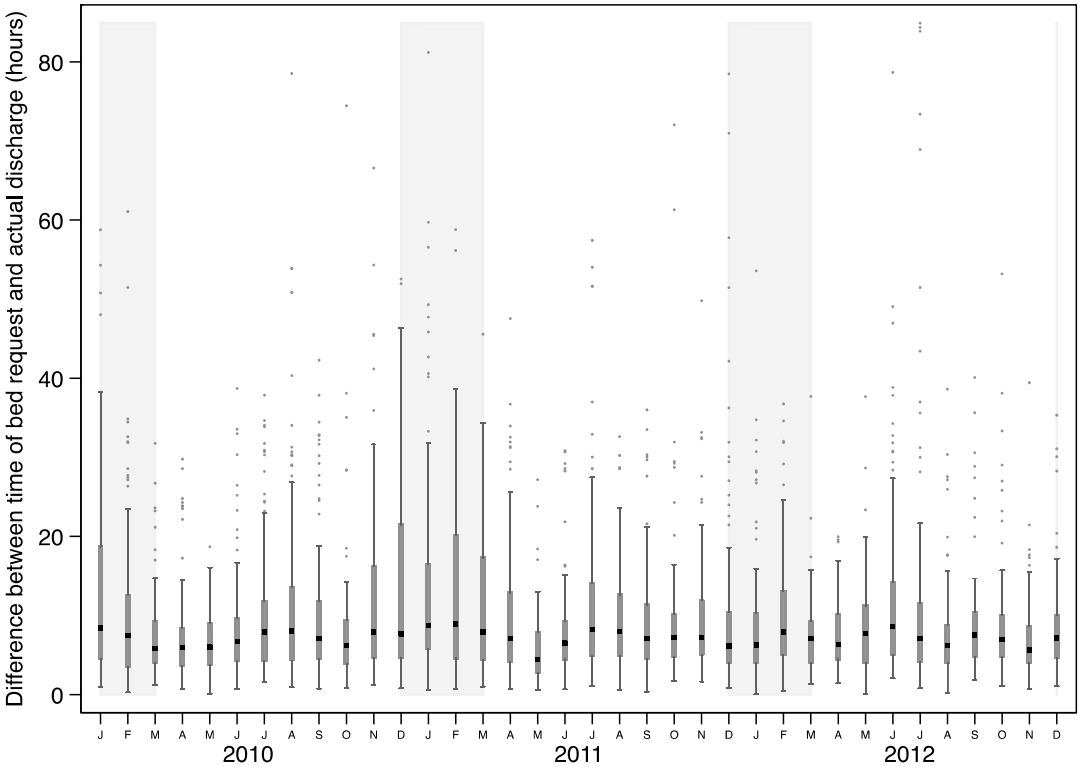


Figure 5.4. Weekly variation of immutable time in the Medical Intensive Care Unit of the Hospital of the University of Pennsylvania, 2010-2012



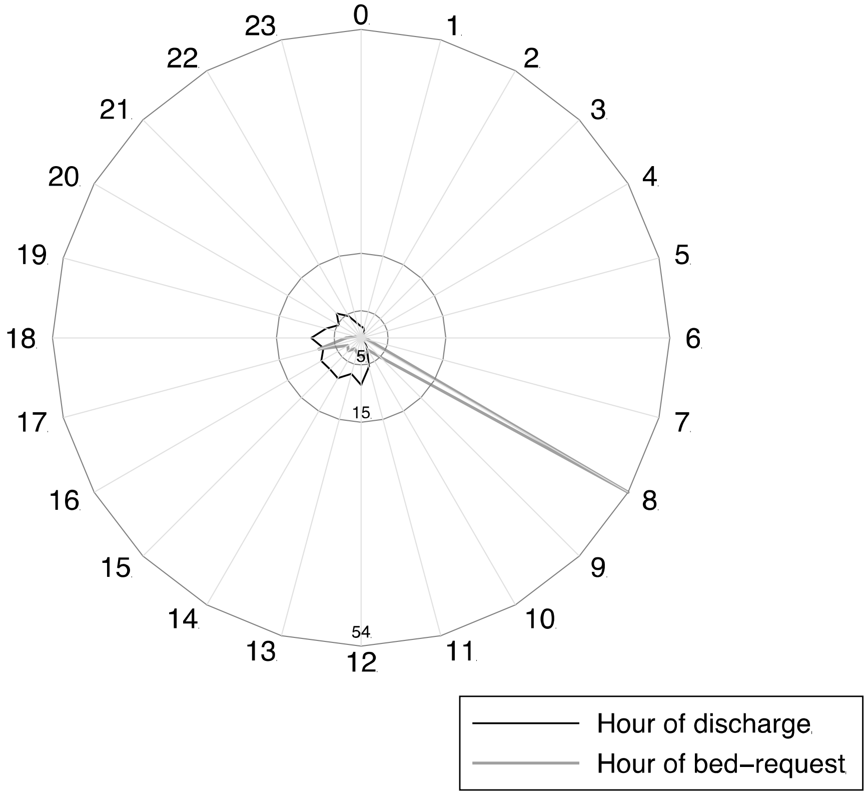
Notes: Weekly (n=156 consecutive weeks starting in January 2010 on the far left through to December 2012 on the far right) distributions (boxplots) are calculated using administrative data from electronic medical records. The figure summarizes 3,851 medical intensive care unit (MICU) discharges, some of which are readmissions (i.e., one patient may contribute >1 discharges). Black lines in the middle of the box indicate the weekly median value. The bottom and top of the box represent the weekly interquartile range (IQR) (i.e., first and third quartile), respectively, and the top and bottom of the whiskers extending from the box represent the largest or lowest value not greater or lower than the IQR times 1.5 for each week. Dots above a weekly distribution indicate a time longer than 1.5 times the IQR for that week. Gray shading indicates weeks during peak flu activity in December, January, February and March according to the Centers for Disease Control.

Figure 5.5. Variation of immutable time by month of the year in the Medical Intensive Care Unit of the Hospital of the University of Pennsylvania, 2010-2012



Notes: Monthly (n=36 consecutive months starting in January 2010 on the far left through to December 2012 on the far right) distributions (boxplots) are calculated using administrative data from electronic medical records. The figure summarizes 3,851 medical intensive care unit (MICU) discharges, some of which are readmissions (i.e., one patient may contribute >1 discharges). Black lines in the middle of the box indicate the monthly median value. The bottom and top of the box represent the monthly interquartile range (IQR) (i.e., first and third quartile), respectively, and the top and bottom of the whiskers extending from the box represent the largest or lowest value not greater or lower than the IQR x 1.5 for each month. Dots above a monthly distribution indicate a time longer than 1.5 times the IQR for that month. Gray shading indicates weeks during peak flu activity in December, January, February and March according to the Centers for Disease Control.

Figure 5.6. Cumulative radar frequency graph of the time of day that a request for discharge was submitted and occurred (24-hour day)



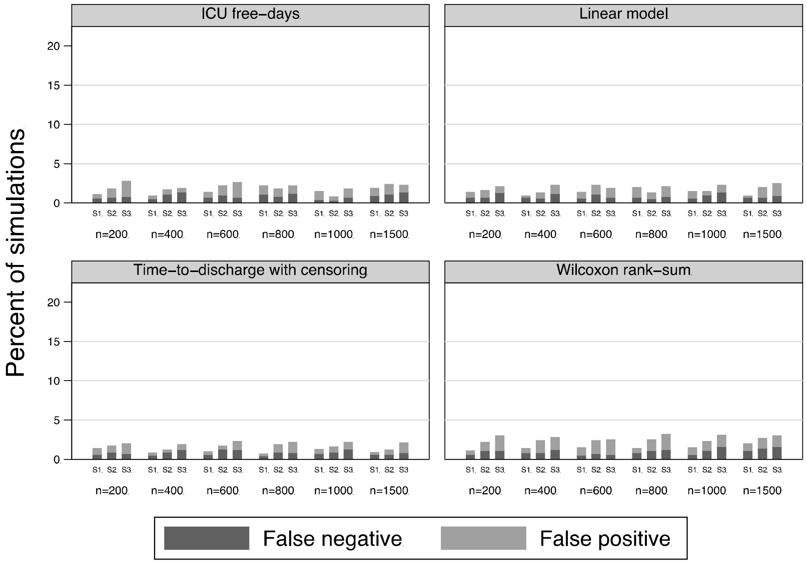
Notes: The figure summarizes the time (using a 24-hour day where 0 is midnight and 23 is 11 pm) a discharge order was placed for 3,851 discharges in the calendar years of 2010 to 2012 (n=156 weeks). The numbers inside the intermediate circles indicate 5%, 15%, and 54% of the total, with the latter representing 2,076 discharges requested between 8-9 am, when morning bed management rounds occur. The second spike of the hour-of-bed-request curve, at roughly 5pm, coincides with the typical timing of evening bed management rounds.

Simulation study

When data were simulated with no treatment effect, with a few exceptions in the largest immutable time setting (i.e., setting 3), less than 2-3% of replicates had an inferential mismatch due to immutable time (Figure 5.7). In the presence of a treatment effect, the rate of inferential mismatches varied considerably depending on the statistical model. Generally, inferential differences between the unbiased and immutable time LOS tended to decrease as (i) sample size increased, (ii) mortality rates decreased, and/or (iii) the magnitude of the treatment effect increased relative to the median immutable time (Figures 5.7-5.13). For example, in settings that simulated a half-day median reduction in LOS with a 20% mortality rate, false inference rates as high as 5% in setting 1 and 15% in setting 3 were observed. In the settings with a full day LOS treatment effect at the median, false inferences tended to be isolated to the smaller sample sizes of 200 to 600. Inferential differences between the two data generation approaches became apparent as mortality increased. The rank-sum test and linear regression models tended to report mostly false-negatives and have higher overall false inferential rates under the competing risks framework. The ICU free-day metric, which valued LOS as 0 for decedents, seemed to reduce inferential errors as the mortality rate increased under the competing risks framework, but less so under the principal stratification framework. The Cox time-to-event model exhibited higher, and mostly false-positive, rates as overall mortality increased from 10% to 40% under principal stratification data generation. These rates were much lower and isolated to small sample sizes in the competing risks framework. In many cases, the false positive and false negative interpretations result from small changes in the observed p-values relative to those that would have been obtained in the absence of immutable time.

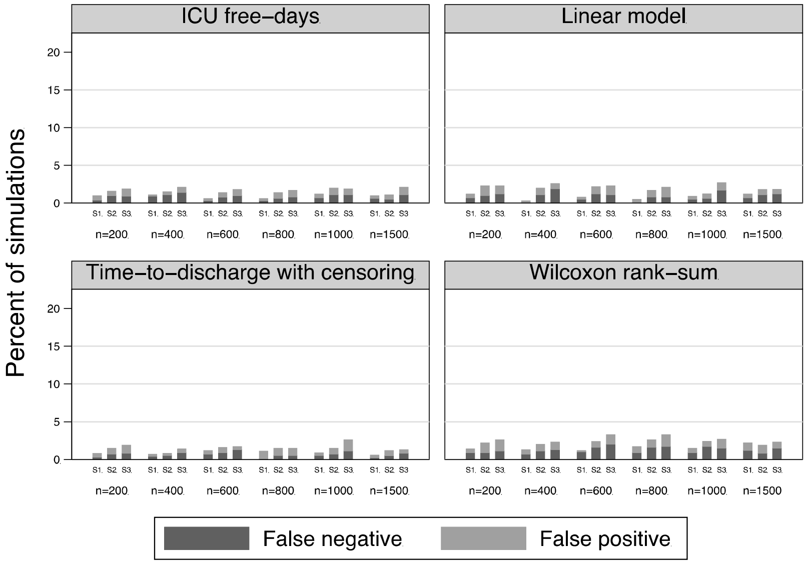
Figure 5.7. Simulation results with no LOS reduction

A) 10% overall mortality rate



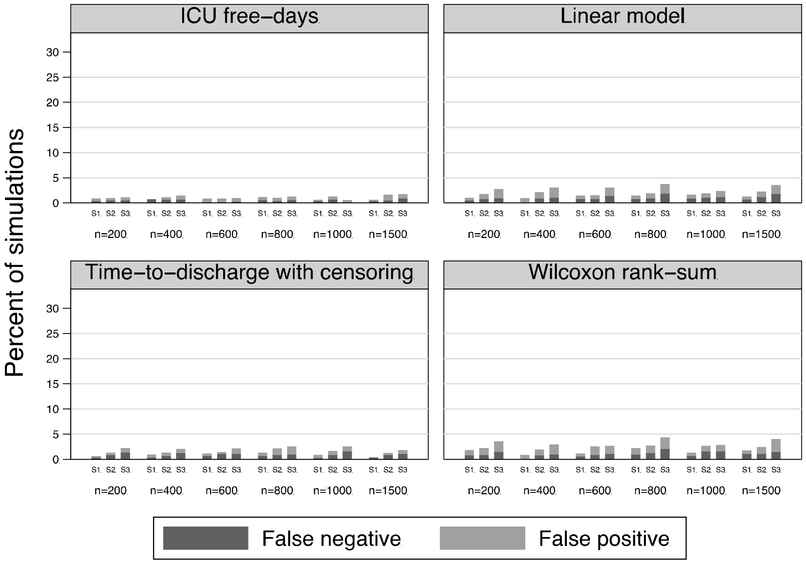
Control group mortality 10%, no LOS reduction
Data generated with a no effect.

B) 20% overall mortality rate



Control group mortality 20%, no LOS reduction
Data generated with a no effect.

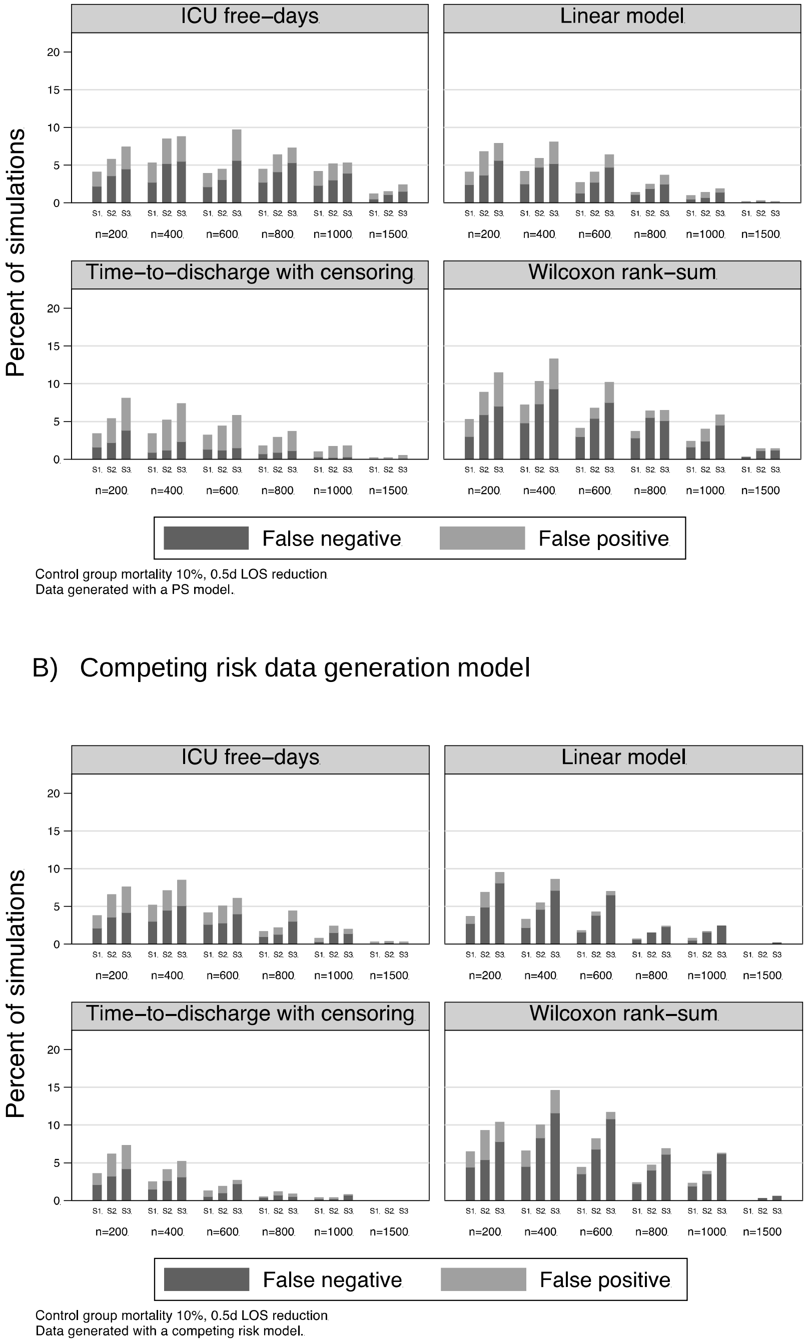
C) 40% overall mortality rate



Control group mortality 40%, no LOS reduction
Data generated with a no effect.

Notes: S1, S2, and S3 indicate immutable time settings 1, 2, and 3, respectively, and correspond with a median of approximately 8, 16, and 28 hours of extra immutable time (Figure 5.2). A false positive is operationally defined for this study as a two-sided statistical test on the difference between study arms with immutable time added finding a p-value <0.05 when the error-free LOS distribution had a p-value ≥0.05. A false negative is operationally defined for this study as a two-sided statistical test on the difference between study arms with immutable time added finding a p-value ≥0.05 when the error-free LOS distribution had a p-value <0.05.

Figure 5.8. Simulation results with 10% baseline mortality, 0.5 day LOS reduction at the median



A) Principal stratification data generation model

Control group mortality 10%, 0.5d LOS reduction
Data generated with a PS model.

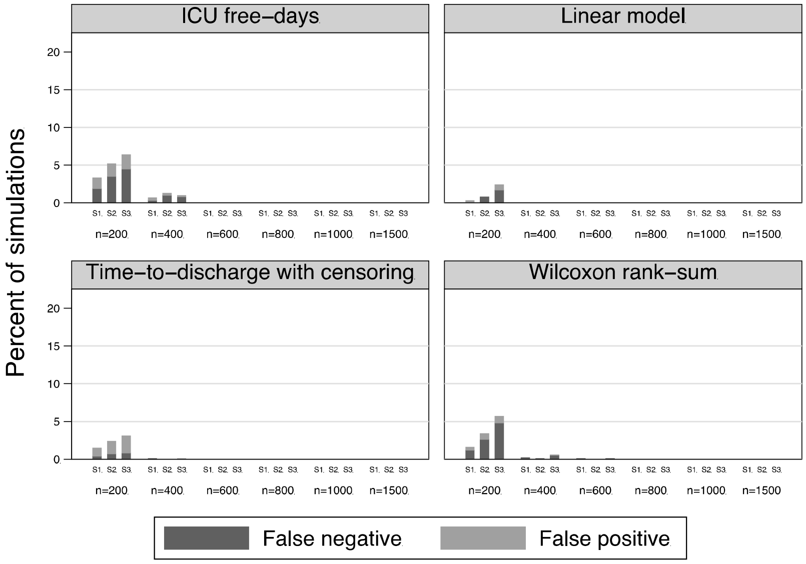
B) Competing risk data generation model

Control group mortality 10%, 0.5d LOS reduction
Data generated with a competing risk model.

Notes: See footnote for Figure 5.7. Reported LOS reduction is at the median.

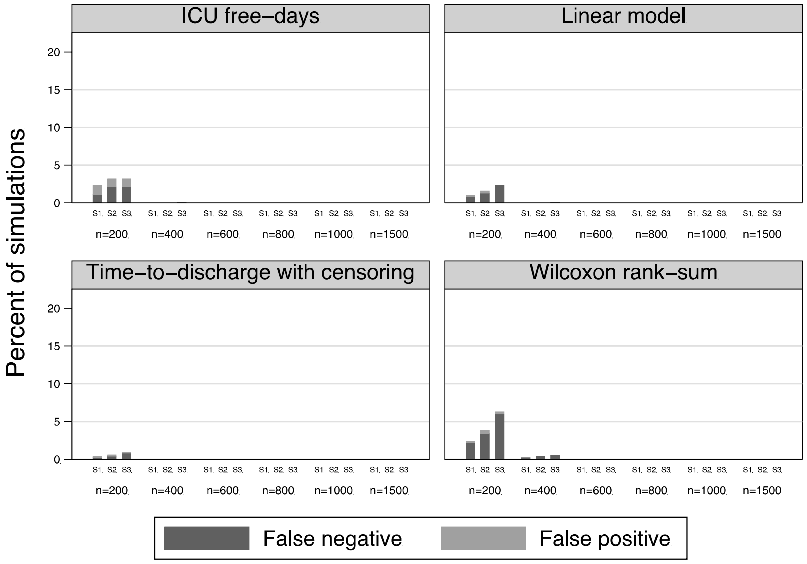
Figure 5.9. Simulation results with 10% baseline mortality, 1 day LOS reduction at the median

A) Principal stratification data generation model



Control group mortality 10%, 1d LOS reduction
Data generated with a PS model.

B) Competing risk data generation model

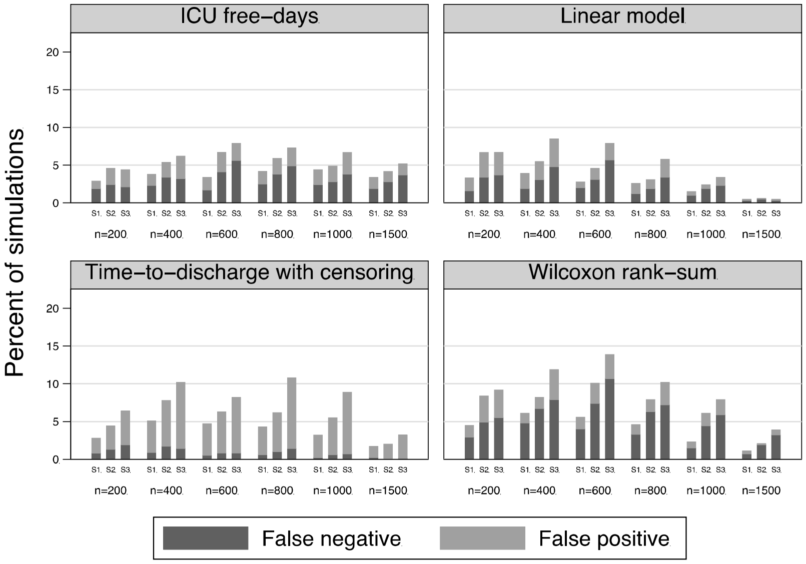


Control group mortality 10%, 1d LOS reduction
Data generated with a competing risk model.

Notes: See footnote for Figure 5.7. Reported LOS reduction is at the median.

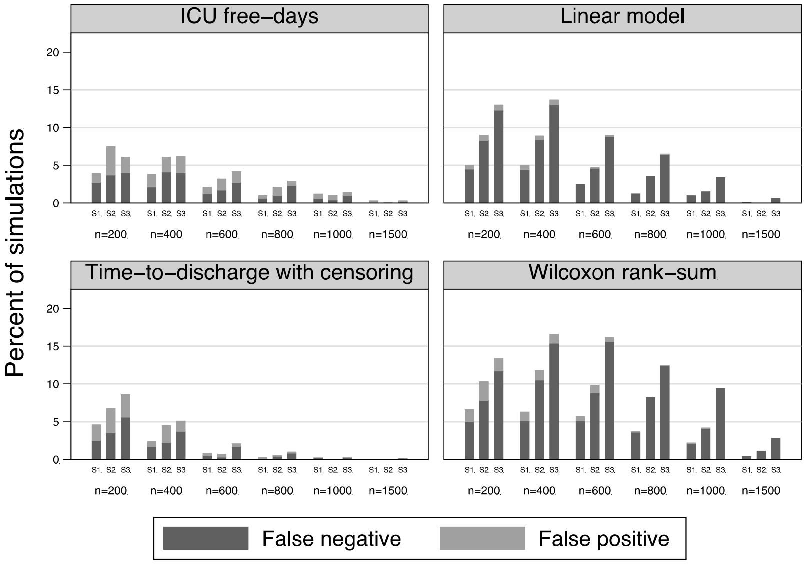
Figure 5.10. Simulation results with 20% baseline mortality, 0.5 day LOS reduction at the median

A) Principal stratification data generation model



Control group mortality 20%, 0.5d LOS reduction
Data generated with a PS model.

B) Competing risk data generation model

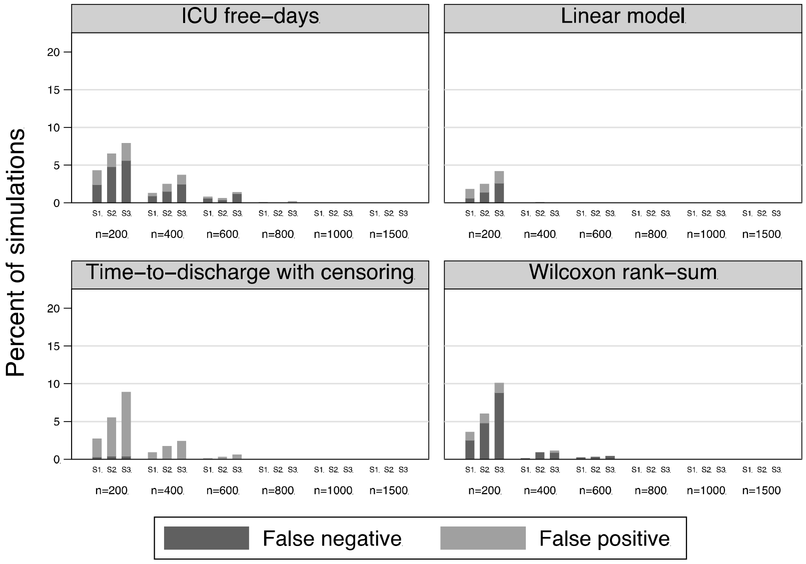


Control group mortality 20%, 0.5d LOS reduction
Data generated with a competing risk model.

Notes: See footnote for Figure 5.7. Reported LOS reduction is at the median.

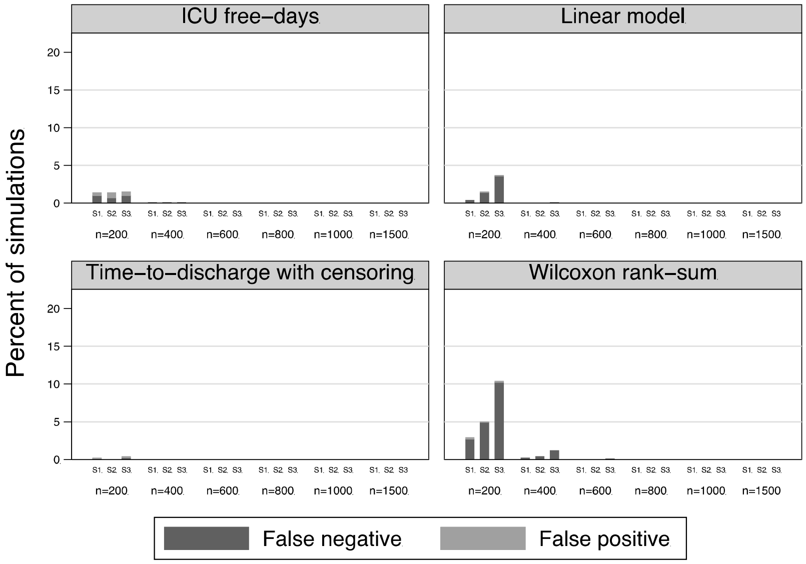
Figure 5.11. Simulation results with 20% baseline mortality, 1 day LOS reduction at the median

A) Principal stratification data generation model



Control group mortality 20%, 1d LOS reduction
Data generated with a PS model.

B) Competing risk data generation model

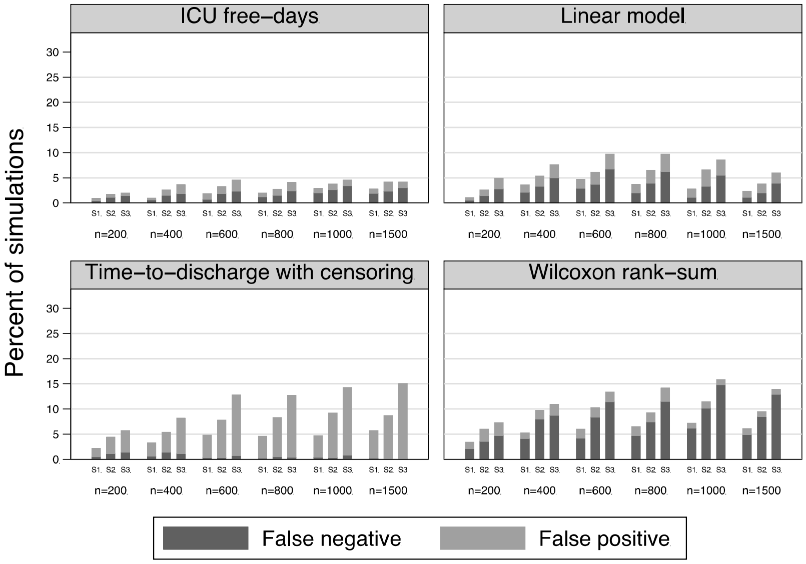


Control group mortality 20%, 1d LOS reduction
Data generated with a competing risk model.

Notes: See footnote for Figure 5.7. Reported LOS reduction is at the median.

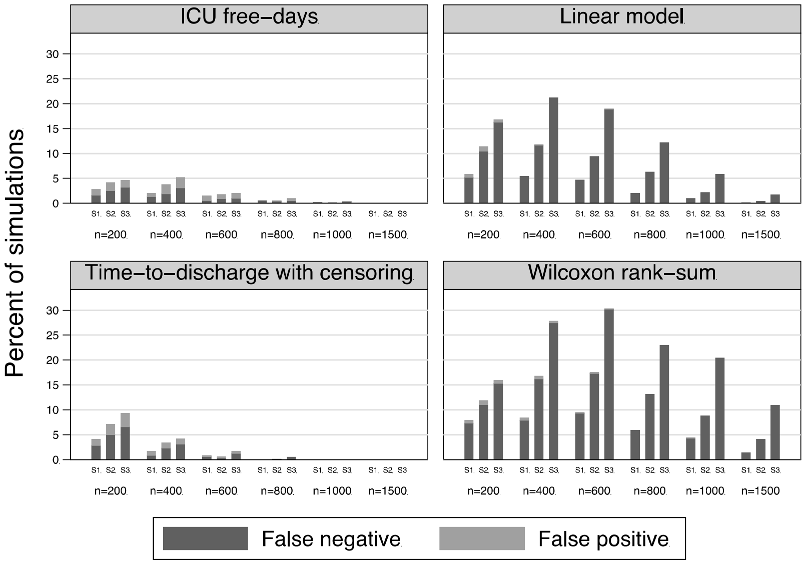
Figure 5.12. Simulation results with 40% baseline mortality, 0.5 day LOS reduction at the median

A) Principal stratification data generation model



Control group mortality 40%, 0.5d LOS reduction
Data generated with a PS model.

B) Competing risk data generation model

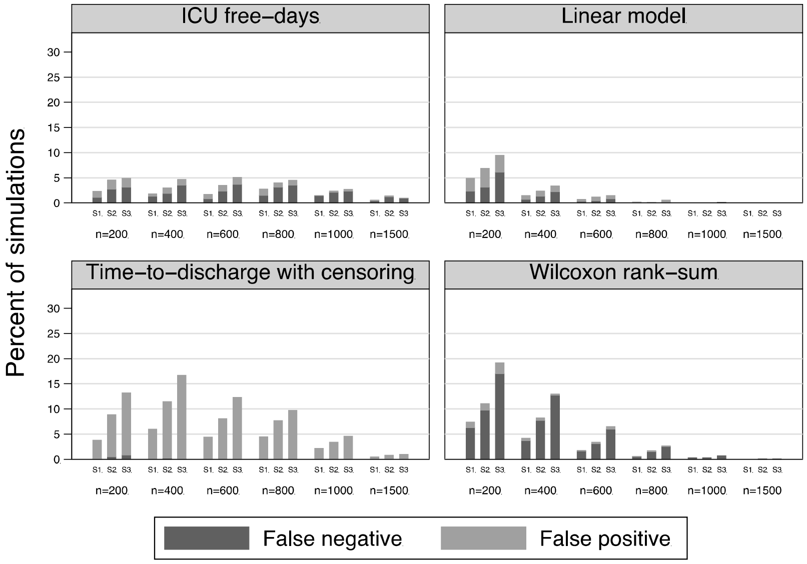


Control group mortality 40%, 0.5d LOS reduction
Data generated with a competing risk model.

Notes: See footnote for Figure 5.7. Reported LOS reduction is at the median.

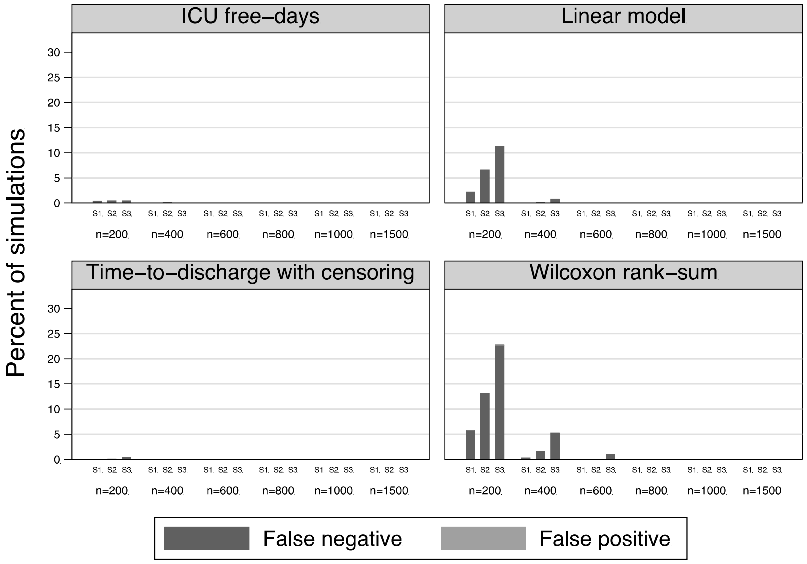
Figure 5.13. Simulation results with 40% baseline mortality, 1 day LOS reduction at the median

A) Principal stratification data generation model



Control group mortality 40%, 1d LOS reduction
Data generated with a PS model.

B) Competing risk data generation model



Control group mortality 40%, 1d LOS reduction
Data generated with a competing risk model.

Notes: See footnote for Figure 5.7. Reported LOS reduction is at the median.

Discussion

The increasing uptake of duration-based endpoints including ICU LOS in health services research requires scrutiny of the various definitions, analytic approaches, and inherent biases of these metrics. The present study provides several important findings regarding the awareness, patterns, and measurement errors associated with ICU LOS. First, fewer than 5% of published ICU-based RCTs in the modern era have reported using either a medical or administrative ready-to-go definition of ICU LOS. Further, there is often a lack of specificity in the reporting of how ICU LOS was measured or analyzed. Without details regarding how authors define LOS endpoints, comparisons of results across trials and meta-analyses are intrinsically limited.

Second, we re-analyzed data from the SUNSET trial and found that participants experienced a median of 5.1 additional hours (IQR 2.7 to 8.9 hours) of ICU time after their clinical improvement was confirmed with a discharge order. This discharge immutable time varied widely in a subsequent examination of our center's administrative data over a three-year period where we found a slightly longer median immutable time of 7.0 hours (IQR 4.3 to 11.1 hours).

Third, we found that even when simplifying the mechanics of ICU LOS in a simulation setting, the addition of immutable time to hypothetical error-free LOS distributions can unpredictably erode statistical inference regardless of the statistical model. In reality ICU LOS is not a homogenous distribution, but rather one comprised of time to death and time to discharge sub-distributions, with heterogeneous patient subgroups (and thus LOS) in both. An intervention can impose complex changes to these various sub-distributions upon which immutable time adds an additional layer of inferential complexity. In our controlled simulation settings, the principal findings are that immutable time could lead to a wrong conclusion about the effect of an intervention on time-to-discharge readiness. While this predominantly results in the dilution or masking of a statistically significant decline in the simulated time-to-discharge readiness when the total LOS was assessed, particularly with small sample sizes, immutable time could also result in instances where the total LOS exhibited statistically significant differences not occurring in time-to-discharge readiness. However, this was predominantly in the principal stratification framework and the time-to-event model and likely attributable to how this model estimates the latent LOS of censored individuals. Nevertheless, this suggests that non-differential measurement error may be less predictable when both the underlying error-free and measurement error distributions are not normally distributed as their combination can result in quantities that are dissimilar in terms of mean, median, and variance. The resulting small changes in the p-values between the two models can become magnified when statistical significance thresholds (i.e., 0.05) are imposed. This has important implications for how interventions are evaluated in trials versus real-world settings. First, it suggests that more than one modeling approach can better inform LOS comparisons. Second, declines in time-to-discharge readiness may be important to determine the efficacy of a treatment but different stakeholders may interpret small declines in this time that don’t result in declines in total ICU LOS differently.

In interpreting this work, the implications of several modeling assumptions must be considered. First, though useful for illustration, an important limitation of our simulation study is the assumption that immutable time would be non-differentially added across arms. This assumption may not be true in practice. For instance, in our analysis of three years of administrative data from one MICU, we found that the duration of discharge immutable time varied by week and year (Figures 5.4 and 5.5), suggesting immutable time may be differentially added to patients treated in the same ICU over time. While this variation would likely become evenly distributed between intervention arms over a sufficiently long recruitment period, particularly if randomization was done within center, these variations could conceivably lead to differential effects of immutable time across arms in some trials. Such differential effects could further distort treatment effect estimates (Fuller, 2006; Hyslop & Imbens, 2001), unless center and seasonal effects are assessed and accounted for analytically (Kahan & Harhay, 2015).

A second assumption that may not hold in clinical practice is that the effects of immutable time would accrue consistently across patient subgroups. In fact, special patient populations may experience levels of immutable time that exceed the average for a given ICU (e.g., patients requiring contact precautions, telemetry, or higher-level nursing observation may take longer to discharge to appropriate ward beds). Thus, trials that accrue relatively larger proportions of such patients may experience larger effects of immutable time on precision.

Third, setting 2 is potentially the most informative to consider how immutable time may manifest in reality and while the third simulation setting can be informative for understanding the impact of large immutable times, such times may not be observed in many real-world settings. In trials there will be a point at which patients are randomized and enrolled (which may not always be equivalent and may or may not be at the time of admission). Thus, while discharge immutable time will almost always exist to some extent, the existence and size of early immutable time will vary with different study designs and interventions or it may be directly related to inclusion criteria (e.g., patients ventilated >48hours). Thus, there are two potential definitions of early immutable time: (1) time in the ICU from admission, or (2) after meeting exclusion criteria but prior to exposure to an intervention that should be considered when designing and analyzing a study.

Finally, the ready-to-go time at our center was commonly observed during predictable times of day, immediately following morning and evening bed management rounds (Figure 5.6). Although many ICUs employ similar strategies for reviewing admission and discharge priorities at discrete times of the day, these patterns could limit the applicability of our simulation data to other units. Specifically, although all patients declared “ready to go” at these bed management rounds were truly eligible for discharge, having been granted that status by their physicians, some of these patients were likely ready to go hours beforehand. Trials conducted in ICUs that more frequently assess patients for discharge readiness may experience either longer or shorter immutable time distributions (with heightened or reduced implications for statistical precision, respectively). Such ICUs could experience longer measured immutable time if ward discharge times are relatively constrained, in which case the longer gaps between the more accurately measured ready-to-go times and actual discharge times would exacerbate immutable time. By contrast, if more frequent assessment enabled more efficient discharge practices, this would limit the generation of immutable time.

There are some additional limitations to our analysis. First, other settings not incorporated into the simulation study, such as longer LOS or impacts of an intervention on time-to-death sub-distribution of ICU LOS, may also impact LOS interpretation. Variations in the precision and detail of reporting by authors as well as differential standards between journals may have caused a small underestimation of measuring time-to-discharge readiness in practice in the systematic review. For instance, we found published trial reports that used a standardized criterion for ICU discharge, but appeared to measure ICU LOS through till discharge. Utilization of standardized discharge criteria suggests awareness of immutable time bias, but it does not fully remove it. More broadly, lack of clarity and insufficient adherence to the CONSORT transparent reporting of trials statement standards of outcome definitions such that they are able to be replicated is an important finding of this work, but also a limitation, highlighting the importance for improved standardization and reporting of ICU LOS and other common outcomes.

In summary, LOS is an attractive endpoint for use in a wide range of healthcare outcomes research because it is important to patients, families, and health systems and readily quantifiable. Indeed, it is the most widely used secondary endpoint in ICU-based RCTs (Harhay et al., 2014). While our study focused on LOS measurement error in the ICU, the immutable time we identified is unlikely to be limited to ICU settings. Therefore, our results have potentially broad significance and applicability for health services research. Thus, when utilizing duration endpoints such as LOS, failing to consider or report definitions of the outcomes that are most plausibly related to interventions may result in inconsistency across trials, reduced power, and potentially even bias. These problems may be especially important in ICU-based RCTs, given the difficulties of recruiting adequate sample sizes to identify realistic and clinically meaningful treatment effects (Aberegg et al., 2010; Harhay et al., 2014; Rubenfeld, 2015).

CHAPTER 6. CONCLUSIONS

There are several overarching conclusions that can be derived from this thematically linked work examining endpoints in ICU-based RCTs. First, our extensive review of the literature demonstrates a lack of standardized definitions and methods for reporting and comparing both mortal (Chapter 2) and nonmortal (Chapters 4 and 5) endpoints in published, experiential critical care research. This finding has several practical consequences. Primarily, as a result of a lack of core and standardized outcomes, critical care researchers who are interested in examining the same endpoint may in fact be asking different empirical questions, especially in regard to endpoints measured in terms of duration. The result manifests in how trials are conducted and reported and thus translated into clinical practice. It also confounds the ability of researches attempting to confirm the external validity of RCT findings, understand differences in observed effects between RCTs, and conducting meta-analyses. A second key finding of this work is that trials of mortality infrequently show a mortality effect due to their sample size (Chapter 2), but the small mortality differences that are observed have the potential to impact how nonmortal endpoints are interpreted (Chapters 4 and 5). This highlights an important fact: *non-significant results do not indicate null results*. We have shown that in simplified simulated settings these issues can profoundly complicate the interpretation of trial results. In real-world settings when all these issues are simultaneously occurring, the inferential complexity is likely intensified and requires close examination of several factors to fully understand how an intervention has impacted patients (Tables 4.1 and 4.2)

In the course of this work, we have identified several new questions that may serve as the foundation for future research. For instance, though trials may adjust for prognostic variables such as severity-of-illness to assess their outcomes, other sources of variation (i.e., treating physician, ICU and hospital factors) are often not accounted for in statistical analysis. This practice has consequences, particularly given the frequent recommendation that more patients and centers should be included in ICU-based RCTs to increase sample sizes and generalizability (Landoni et al., 2015). While the multicenter design increases the likelihood of attaining adequate sample sizes, the diverse characteristics that vary across ICUs, such as intensivist practice patterns, protocols, or the acumen of trainees and staff can exert a powerful influence on patient outcomes. Thus, I contend that primary endpoints should be adjusted, at a minimum, for patient acuity and center effects in primary analyses. This is especially important when the contribution of enrolled subjects differs across centers, and failure to account for center effects can lead to complex and unpredictable type I or II errors when comparing trial arms (Kahan & Harhay, 2015; Kahan et al., 2014; Kahan & Morris, 2013). A key focus of my future research will be on how to correctly account for these potential drivers and confounders of patient outcomes.

A second research direction relates to identification of better endpoints for trials. In Chapter 2, we describe that nonmortal endpoints have become common in ICU RCTs, but in Chapter 4 we illustrate the point that nonmortal endpoints are interlinked with and confounded by mortality. As a potential solution, the use of event-free-days as a composite outcome (whereby death equals zero free-days) is gaining traction (Blackwood et al., 2014; Harhay et al., 2014; National Heart et al., 2011; Rice et al., 2011) because this composite outcome offers more statistical power than mortality endpoints (Schoenfeld et al., 2002) and also overcomes certain biases associated with death-induced missing data (i.e., informative dropout), which can complicate the interpretation of duration-based endpoints (see Chapter 4). However, traditional analyses of free-day outcomes do not describe patient trajectories over time (Schoenfeld et al., 2002) and disregard patient preferences, some of whom may choose death over different long-term care requirements (Rubin et al., 2016). Finally, when interventions influence mortality, independent examination of the mortal and nonmortal endpoints is necessary, thereby limiting the conceptual, inferential and statistical benefits of free-day metrics as an outcome (Cannon, 1997; Freemantle et al., 2003; Schoenfeld et al., 2002; Tomlinson & Detsky, 2010).

There are several potential approaches to meet the need for better trial endpoints. The first is to develop and validate a weighting within a composite outcome framework that accounts for patients' preferences for death over different disease and treatment states. The second approach relates to the assessment of endpoints with more appropriate statistical methods. In this regard I am especially interested and actively pursuing research to study nonmortal outcomes such as ICU free-days, fluid balance and sequential organ failure assessment (SOFA) scores in a joint longitudinal and time-to-event framework. Joint longitudinal and survival modeling methods uniquely permit the simultaneous assessment of the longitudinal nonmortal endpoint (e.g., free-days over time) and the mortal endpoint in a single statistical model (Ratcliffe et al., 2004). This approach represents a novel analytic method to incorporate data on informative dropout that generates unbiased treatment effects and provides greater insight regarding patient trajectories over time (Ratcliffe et al., 2004; Rizopoulos, 2012). Integrating a patient-centered outcome into this longitudinal framework has great promise, as longitudinal statistical methods wield much greater power, thereby potentially reducing needed sample sizes such that resources can be redirected to improve and expand outcome measurement and follow-up. Research by my colleagues has already broached the methodological (Lin et al., 2014) and patient-centered elements (Rubin et al., 2016) of this important topic. Integrating patient-centered outcomes and longitudinal data structures is a necessary and logical next step to make advances in the development of meaningful and measurable patient-centered outcomes in critical care.

Concluding remarks

In conclusion, this dissertation consisted of three empirical analyses that focused on describing the current landscape of ICU-based RCTs and illustrating the consequences of bias and measurement error in the interpretation of trial results. Chapter 2 established the foundation upon which Chapters 4 and 5 were built. Accordingly, the conclusions and recommendations of Chapter 2 are even more relevant when the work is considered together. Dialogue and agreement throughout the critical care community (including journal editors) that goes beyond another “call-to-action” is needed. Indeed, a subspecialty-specific CONSORT guideline statement may be warranted given the unique challenges identified and number of non-significant trials that have been published. The International Forum for Acute Care Trialists (InFACT) is actively pursuing similar endeavors on a smaller scale. Their work, for instance, is currently focusing on the identification and publication of a minimum core set of outcomes for ventilation and other modality-specific interventions (Blackwood et al., 2014; Blackwood et al., 2015). Such a forum would be most appropriate for developing a critical care CONSORT document. This dissertation also highlights the need for methodological work to improve strategies for defining and measuring nonmortal outcomes, in such a manner that trial results can be compared across numerous centers and patient populations. To this end, the results from this dissertation and our recommendations will hopefully have a positive, albeit incremental, impact on the future of trials in critical care patients.

BIBLIOGRAPHY

Aalen, O.O., Cook, R.J., & Roysland, K. (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Anal,* 21, 579-593.

Aberegg, S.K., Richards, D.R., & O'Brien, J.M. (2010). Delta inflation: a bias in the design of randomized controlled trials in critical care medicine. *Crit Care,* 14, R77.

Adhikari, N.K., Tansey, C.M., McAndrews, M.P., Matte, A., Pinto, R., Cheung, A.M., et al. (2011). Self-reported depressive symptoms and memory complaints in survivors five years after ARDS. *Chest,* 140, 1484-1493.

Ali, N.A., Hammersley, J., Hoffmann, S.P., O'Brien, J.M., Jr., Phillips, G.S., Rashkin, M., et al. (2011). Continuity of care in intensive care units: a cluster-randomized trial of intensivist staffing. *Am J Respir Crit Care Med,* 184, 803-808.

Altman, D.G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ,* 332, 1080.

Amrein, K., Schnedl, C., Holl, A., Riedl, R., Christopher, K.B., Pachler, C., et al. (2014). Effect of high-dose vitamin D3 on hospital length of stay in critically ill patients with vitamin D deficiency: the VITdAL-ICU randomized clinical trial. *JAMA,* 312, 1520-1530.

Angus, D.C., Barnato, A.E., Linde-Zwirble, W.T., Weissfeld, L.A., Watson, R.S., Rickert, T., et al. (2004). Use of intensive care at the end of life in the United States: an epidemiologic study. *Critical Care Medicine,* 32, 638-643.

Angus, D.C., Mira, J.P., & Vincent, J.L. (2010). Improving clinical trials in the critically ill. *Critical Care Medicine,* 38, 527-532.

Angus, D.C., & van der Poll, T. (2013). Severe sepsis and septic shock. *N Engl J Med,* 369, 840-851.

Annane, D. (2009). Improving clinical trials in the critically ill: unique challenge--sepsis. *Critical Care Medicine,* 37, S117-128.

ARDSnet Investigators (2000). Ventilation with Lower Tidal Volumes as Compared with Traditional Tidal Volumes for Acute Lung Injury and the Acute Respiratory Distress Syndrome. *New England Journal of Medicine,* 342, 1301-1308.

Azoulay, E., Pochard, F., Kentish-Barnes, N., Chevret, S., Aboab, J., Adrie, C., et al. (2005). Risk of post-traumatic stress symptoms in family members of intensive care unit patients. *American Journal of Respiratory and Critical Care Medicine,* 171, 987-994.

Beyersmann, J., Latouche, A., Buchholz, A., & Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Stat Med,* 28, 956-971.

Bienvenu, O.J., Colantuoni, E., Mendez-Tellez, P.A., Dinglas, V.D., Shanholtz, C., Husain, N., et al. (2012). Depressive symptoms and impaired physical function after acute lung injury: a 2-year longitudinal study. *Am J Respir Crit Care Med,* 185, 517-524.

Blackwood, B., Clarke, M., McAuley, D.F., McGuigan, P.J., Marshall, J.C., & Rose, L. (2014). How outcomes are defined in clinical trials of mechanically ventilated adults and children. *Am J Respir Crit Care Med,* 189, 886-893.

Blackwood, B., Marshall, J., & Rose, L. (2015). Progress on core outcome sets for critical care research. *Curr Opin Crit Care,* 21, 439-444.

Brock, G.N., Barnes, C., Ramirez, J.A., & Myers, J. (2011). How to handle mortality when investigating length of hospital stay and time to clinical stability. *BMC Med Res Methodol,* 11, 144.

Cameron, J.I., Chu, L.M., Matte, A., Tomlinson, G., Chan, L., Thomas, C., et al. (2016). One-Year Outcomes in Caregivers of Critically Ill Patients. *N Engl J Med,* 374, 1831-1841.

Cannon, C.P. (1997). Clinical perspectives on the use of composite endpoints. *Controlled Clinical Trials,* 18, 517-529; discussion 546-519.

Casaer, M.P., Mesotten, D., Hermans, G., Wouters, P.J., Schetz, M., Meyfroidt, G., et al. (2011). Early versus late parenteral nutrition in critically ill adults. *N Engl J Med,* 365, 506-517.

Checkley, W., Brower, R.G., Munoz, A., & Investigators, N.I.H.A.R.D.S.N. (2010). Inference for mutually exclusive competing events through a mixture of generalized gamma distributions. *Epidemiology,* 21, 557-565.

Chiba, Y., & VanderWeele, T.J. (2011). A simple method for principal strata effects when the outcome has been truncated due to death. *Am J Epidemiol,* 173, 745-751.

Chiche, J.D., & Angus, D.C. (2008). Testing protocols in the intensive care unit: complex trials of complex interventions for complex patients. *JAMA : the journal of the American Medical Association,* 299, 693-695.

Contentin, L., Ehrmann, S., & Giraudeau, B. (2014). Heterogeneity in the definition of mechanical ventilation duration and ventilator-free days. *Am J Respir Crit Care Med,* 189, 998-1002.

Cook, D., & Rocker, G. (2014). Dying with dignity in the intensive care unit. *N Engl J Med,* 370, 2506-2514.

Cooke, C.R. (2012). Economics of mechanical ventilation and respiratory failure. *Crit Care Clin,* 28, 39-55, vi.

Cox, C.E., Docherty, S.L., Brandon, D.H., Whaley, C., Attix, D.K., Clay, A.S., et al. (2009). Surviving critical illness: acute respiratory distress syndrome as experienced by patients and their caregivers. *Crit Care Med,* 37, 2702-2708.

Cox, D.R., Fitzpatrick, R., Fletcher, A.E., Gore, S.M., Spiegelhalter, D.J., & Jones, D.R. (1992). Quality-of-life assessment: Can we keep it simple? *Journal of the Royal Statistical Society: Series A,* 155, 353-393.

Crowther, M.J. (2011). SURVSIM: Stata Module to Simulate Complex Survival Data, Statistical Software Components. Boston College Department of Economics: Boston. Available from: http://ideas.repec.org/c/boc/bocode/s457317html.

Crowther, M.J., & Lambert, P.C. (2012). Simulating complex survival data. *Stata Journal,* 12, 674–687.

Dasta, J.F., McLaughlin, T.P., Mody, S.H., & Piech, C.T. (2005). Daily cost of an intensive care unit day: the contribution of mechanical ventilation. *Crit Care Med,* 33, 1266-1271.

Deddens, J.A., & Petersen, M.R. (2008). Approaches for estimating prevalence ratios. *Occup Environ Med,* 65, 481, 501-486.

Deslandes, E., & Chevret, S. (2010). Joint modeling of multivariate longitudinal data and the dropout process in a competing risk setting: application to ICU data. *BMC Med Res Methodol,* 10, 69.

Dreyfuss, D. (2004). Beyond randomized, controlled trials *Current Opinion in Critical Care,* 10, 574-578.

Fan, E., Dowdy, D.W., Colantuoni, E., Mendez-Tellez, P.A., Sevransky, J.E., Shanholtz, C., et al. (2014). Physical complications in acute lung injury survivors: a two-year longitudinal prospective study. *Crit Care Med,* 42, 849-859.

Ferguson, N.D., Scales, D.C., Pinto, R., Wilcox, M.E., Cook, D.J., Guyatt, G.H., et al. (2013). Integrating Mortality and Morbidity Outcomes: Using Quality-adjusted Life Years in Critical Care Trials. *American Journal of Respiratory and Critical Care Medicine,* 187, 256-261.

Fleischmann, C., Scherag, A., Adhikari, N.K., Hartog, C.S., Tsaganos, T., Schlattmann, P., et al. (2016). Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. Current Estimates and Limitations. *Am J Respir Crit Care Med,* 193, 259-272.

Force, A.D.T., Ranieri, V.M., Rubenfeld, G.D., Thompson, B.T., Ferguson, N.D., Caldwell, E., et al. (2012). Acute respiratory distress syndrome: the Berlin Definition. *JAMA,* 307, 2526-2533.

Freemantle, N., Calvert, M., Wood, J., Eastaugh, J., & Griffin, C. (2003). Composite outcomes in randomized trials: greater precision but with greater uncertainty? *JAMA : the journal of the American Medical Association,* 289, 2554-2559.

Friede, T., & Kieser, M. (2006). Sample size recalculation in internal pilot study designs: a review. *Biom J,* 48, 537-555.

Fuller, W.A. (2006). *Measurement error models*. Hoboken, N.J.: Wiley-Interscience.

Garland, A., & Connors, A.F. (2007). Physicians' influence over decisions to forego life support. *J Palliat Med,* 10, 1298-1305.

Girard, T.D., Kress, J.P., Fuchs, B.D., Thomason, J.W., Schweickert, W.D., Pun, B.T., et al. (2008). Efficacy and safety of a paired sedation and ventilator weaning protocol for mechanically ventilated patients in intensive care (Awakening and Breathing Controlled trial): a randomised controlled trial. *Lancet,* 371, 126-134.

Guerin, C., Reignier, J., Richard, J.C., Beuret, P., Gacouin, A., Boulain, T., et al. (2013). Prone positioning in severe acute respiratory distress syndrome. *N Engl J Med,* 368, 2159-2168.

Halpern, N.A., Goldman, D.A., Tan, K.S., & Pastores, S.M. (2016). Trends in Critical Care Beds and Use Among Population Groups and Medicare and Medicaid Beneficiaries in the United States: 2000-2010. *Crit Care Med,* 44, 1490-1499.

Halpern, N.A., & Pastores, S. (2010). Critical care medicine in the United States 2000-2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Critical Care Medicine,* 38, 65-71.

Halpern, S.D. (2011a). Financial incentives for research participation: empirical questions, available answers and the burden of further proof. *Am J Med Sci,* 342, 290-293.

Halpern, S.D. (2011b). Perceived Inappropriateness of Care in the ICU What to Make of the Clinician's Perspective? *Jama-Journal of the American Medical Association,* 306, 2725-2726.

Halpern, S.D., Karlawish, J.H., & Berlin, J.A. (2002). The continuing unethical conduct of underpowered clinical trials. *JAMA,* 288, 358-362.

Hamric, A.B., & Blackhall, L.J. (2007). Nurse-physician perspectives on the care of dying patients in intensive care units: Collaboration, moral distress and ethical climate. *Crit Care Med,* 35, 422-429.

Harhay, M.O., Wagner, J., Ratcliffe, S.J., Bronheim, R.S., Gopal, A., Green, S., et al. (2014). Outcomes and statistical power in adult critical care randomized trials. *Am J Respir Crit Care Med,* 189, 1469-1478.

Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J.G. (2009). Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform,* 42, 377-381.

Hayden, D., Pauler, D.K., & Schoenfeld, D. (2005). An estimator for treatment comparisons among survivors in randomized trials. *Biometrics,* 61, 305-310.

Hebert, P.C., Wells, G., Blajchman, M.A., Marshall, J., Martin, C., Pagliarello, G., et al. (1999). A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. *New England Journal of Medicine,* 340, 409-417.

Hernán, M.A., & Robins, J.M. (2016). *Causal Inference*.

Hernandez, A.V., Steyerberg, E.W., & Habbema, J.D. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol,* 57, 454-460.

Herridge, M.S., Tansey, C.M., Matte, A., Tomlinson, G., Diaz-Granados, N., Cooper, A., et al. (2011). Functional disability 5 years after acute respiratory distress syndrome. *N Engl J Med,* 364, 1293-1304.

Hirsch, B.R., Califf, R.M., Cheng, S.K., Tasneem, A., Horton, J., Chiswell, K., et al. (2013). Characteristics of oncology clinical trials: insights from a systematic analysis of ClinicalTrials.gov. *JAMA Intern Med,* 173, 972-979.

Holloway, R.G., & Quill, T.E. (2007). Mortality as a measure of quality: implications for palliative and end-of-life care. *JAMA,* 298, 802-804.

Hopewell, S., Clarke, M., Moher, D., Wager, E., Middleton, P., Altman, D.G., et al. (2008). CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS medicine,* 5, e20.

Hyslop, D.R., & Imbens, G.W. (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics,* 19, 475-481.

Iwashyna, T.J. (2010). Survivorship will be the defining challenge of critical care in the 21st century. *Ann Intern Med,* 153, 204-205.

Iwashyna, T.J., Burke, J.F., Sussman, J.B., Prescott, H.C., Hayward, R.A., & Angus, D.C. (2015). Implications of Heterogeneity of Treatment Effect for Reporting and Analysis of Randomized Trials in Critical Care. *Am J Respir Crit Care Med,* 192, 1045-1051.

Jadad, A.R., Moore, R.A., Carroll, D., Jenkinson, C., Reynolds, D.J.M., Gavaghan, D.J., et al. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials,* 17, 1-12.

Jakob, S.M., Ruokonen, E., Grounds, R.M., Sarapohja, T., Garratt, C., Pocock, S.J., et al. (2012). Dexmedetomidine vs midazolam or propofol for sedation during prolonged mechanical ventilation: two randomized controlled trials. *JAMA,* 307, 1151-1160.

Jansen, T.C., van Bommel, J., Schoonderbeek, F.J., Sleeswijk Visser, S.J., van der Klooster, J.M., Lima, A.P., et al. (2010). Early lactate-guided therapy in intensive care unit patients: a multicenter, open-label, randomized controlled trial. *Am J Respir Crit Care Med,* 182, 752-761.

Juni, P., Altman, D.G., & Egger, M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ,* 323, 42-46.

Kahan, B.C., & Harhay, M.O. (2015). Many multicenter trials had few events per center, requiring analysis via random-effects models or GEEs. *J Clin Epidemiol,* 68, 1504-1511.

Kahan, B.C., Jairath, V., Dore, C.J., & Morris, T.P. (2014). The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials,* 15, 139.

Kahan, B.C., & Morris, T.P. (2013). Assessing potential sources of clustering in individually randomised trials. *BMC Med Res Methodol,* 13, 58.

Kahn, J.M., Rubenfeld, G.D., Rohrbach, J., & Fuchs, B.D. (2008). Cost savings attributable to reductions in intensive care unit length of stay for mechanically ventilated patients. *Med Care,* 46, 1226-1233.

Kent, D.M., Alsheikh-Ali, A., & Hayward, R.A. (2008). Competing risk and heterogeneity of treatment effect in clinical trials. *Trials,* 9, 30.

Kent, D.M., & Hayward, R.A. (2007). Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA,* 298, 1209-1212.

Kerlin, M.P., Small, D.S., Cooney, E., Fuchs, B.D., Bellini, L.M., Mikkelsen, M.E., et al. (2013). A randomized trial of nighttime physician staffing in an intensive care unit. *N Engl J Med,* 368, 2201-2209.

Kress, J.P., & Hall, J.B. (2014). ICU-acquired weakness and recovery from critical illness. *N Engl J Med,* 370, 1626-1635.

Kress, J.P., Pohlman, A.S., O'Connor, M.F., & Hall, J.B. (2000). Daily interruption of sedative infusions in critically ill patients undergoing mechanical ventilation. *The New England journal of medicine,* 342, 1471-1477.

Landoni, G., Comis, M., Conte, M., Finco, G., Mucchetti, M., Paternoster, G., et al. (2015). Mortality in Multicenter Critical Care Trials: An Analysis of Interventions With a Significant Effect. *Crit Care Med,* 43, 1559-1568.

Lash, T.L., Fink, A.K., & Fox, M.P. (2009). *Multidimensional bias analysis. In Applying Quantitative Bias Analysis to Epidemiologic Data*. New York: Springer.

Latronico, N., Metelli, M., Turin, M., Piva, S., Rasulo, F.A., & Minelli, C. (2013). Quality of reporting of randomized controlled trials published in Intensive Care Medicine from 2001 to 2010. *Intensive Care Medicine,* 39, 1386-1395.

Lewis, R.J., Viele, K., Broglio, K., Berry, S.M., & Jones, A.E. (2013). An adaptive, phase II, dose-finding clinical trial design to evaluate L-carnitine in the treatment of septic shock based on efficacy and predictive probability of subsequent phase III success. *Crit Care Med,* 41, 1674-1678.

Lin, W., Halpern, S.D., Prasad Kerlin, M., & Small, D.S. (2014). A "placement of death" approach for studies of treatment effects on ICU length of stay. *Stat Methods Med Res*.

Luce, J.M., Cook, D.J., Martin, T.R., Angus, D.C., Boushey, H.A., Curtis, J.R., et al. (2004). The ethical conduct of clinical research involving critically ill patients in the United States and Canada: principles and recommendations. *Am J Respir Crit Care Med,* 170, 1375-1384.

Marini, J.J. (2006). Limitations of clinical trials in acute lung injury and acute respiratory distress syndrome. *Curr Opin Crit Care,* 12, 25-31.

McAuley, D.F., O'Kane, C., & Griffiths, M.J. (2010). A stepwise approach to justify phase III randomized clinical trials and enhance the likelihood of a positive result. *Crit Care Med,* 38, S523-527.

McConnell, S., Stuart, E.A., & Devaney, B. (2008). The truncation-by-death problem: what to do in an experimental evaluation when the outcome is not always defined. *Eval Rev,* 32, 157-186.

McHugh, G.S., Butcher, I., Steyerberg, E.W., Marmarou, A., Lu, J., Lingsma, H.F., et al. (2010). A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. *Clin Trials,* 7, 44-57.

Mebazaa, A., Laterre, P.F., Russell, J.A., Bergmann, A., Gattinoni, L., Gayat, E., et al. (2016). Designing phase 3 sepsis trials: application of learned experiences from critical care trials in acute heart failure. *J Intensive Care,* 4, 24.

Meltzer, L.S., & Huckabay, L.M. (2004). Critical care nurses' perceptions of futile care and its effect on burnout. *American Journal of Critical Care,* 13, 202-208.

Mikkelsen, M.E., Christie, J.D., Lanken, P.N., Biester, R.C., Thompson, B.T., Bellamy, S.L., et al. (2012). The adult respiratory distress syndrome cognitive outcomes study: long-term neuropsychological function in survivors of acute lung injury. *Am J Respir Crit Care Med,* 185, 1307-1315.

Milbrandt, E.B., Kersten, A., Rahim, M., Dremsizov, T.T., Clermont, G., Cooper, L.M., et al. (2008). Growth of intensive care unit resource use and its estimated cost in Medicare. *Critical Care Medicine,* 36, 2504-2510.

Moitra, V.K., Guerra, C., Linde-Zwirble, W.T., & Wunsch, H. (2016). Relationship Between ICU Length of Stay and Long-Term Mortality for Elderly ICU Survivors. *Crit Care Med,* 44, 655-662.

Morgan, S.L., & Winship, C. (2015). *Counterfactuals and causal inference : methods and principles for social research*. New York, NY: Cambridge University Press.

Murthy, S., Leligdowicz, A., & Adhikari, N.K. (2015). Intensive care unit capacity in low-income countries: a systematic review. *PLoS One,* 10, e0116949.

Myburgh, J.A., Finfer, S., Bellomo, R., Billot, L., Cass, A., Gattas, D., et al. (2012). Hydroxyethyl starch or saline for fluid resuscitation in intensive care. *N Engl J Med,* 367, 1901-1911.

National Heart, L., Blood Institute Acute Respiratory Distress Syndrome Clinical Trials, N., Matthay, M.A., Brower, R.G., Carson, S., Douglas, I.S., et al. (2011). Randomized, placebo-controlled clinical trial of an aerosolized beta(2)-agonist for treatment of acute lung injury. *Am J Respir Crit Care Med,* 184, 561-568.

Naylor, C.D., & Llewellyn-Thomas, H.A. (1994). Can there be a more patient-centred approach to determining clinically important effect sizes for randomized treatment trials? *J Clin Epidemiol,* 47, 787-795.

Opal, S.M., Dellinger, R.P., Vincent, J.L., Masur, H., & Angus, D.C. (2014). The next generation of sepsis clinical trial designs: what is next after the demise of recombinant human activated protein C?*. *Crit Care Med,* 42, 1714-1721.

Ospina-Tascon, G.A., Buchele, G.L., & Vincent, J.L. (2008). Multicenter, randomized, controlled trials evaluating mortality in intensive care: doomed to fail? *Crit Care Med,* 36, 1311-1322.

Papazian, L., Forel, J.M., Gacouin, A., Penot-Ragon, C., Perrin, G., Loundou, A., et al. (2010). Neuromuscular blockers in early acute respiratory distress syndrome. *N Engl J Med,* 363, 1107-1116.

Perner, A., Haase, N., Guttormsen, A.B., Tenhunen, J., Klemenzson, G., Aneman, A., et al. (2012). Hydroxyethyl starch 130/0.42 versus Ringer's acetate in severe sepsis. *N Engl J Med,* 367, 124-134.

Piers, R.D., Azoulay, E., Ricou, B., Ganz, F.D., Decruyenaere, J., Max, A., et al. (2011). Perceptions of Appropriateness of Care Among European and Israeli Intensive Care Unit

Nurses and Physicians. *Jama-Journal of the American Medical Association,* 306, 2694-2703.

Prinsen, C.A., Vohra, S., Rose, M.R., King-Jones, S., Ishaque, S., Bhaloo, Z., et al. (2014). Core Outcome Measures in Effectiveness Trials (COMET) initiative: protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set'. *Trials,* 15, 247.

Ranieri, V.M., Thompson, B.T., Barie, P.S., Dhainaut, J.F., Douglas, I.S., Finfer, S., et al. (2012). Drotrecogin alfa (activated) in adults with septic shock. *N Engl J Med,* 366, 2055-2064.

Rapoport, J., Teres, D., Zhao, Y., & Lemeshow, S. (2003). Length of stay data as a guide to hospital economic performance for ICU patients. *Med Care,* 41, 386-397.

Ratcliffe, S.J., Guo, W., & Ten Have, T.R. (2004). Joint modeling of longitudinal and survival data via a common frailty. *Biometrics,* 60, 892-899.

Reade, M.C., & Angus, D.C. (2009). The clinical research enterprise in critical care: what's right, what's wrong, and what's ahead? *Crit Care Med,* 37, S1-9.

Resche-Rigon, M., Azoulay, E., & Chevret, S. (2006). Evaluating mortality in intensive care units: contribution of competing risks analyses. *Crit Care,* 10, R5.

Rice, T.W., Wheeler, A.P., Thompson, B.T., deBoisblanc, B.P., Steingrub, J., Rock, P., et al. (2011). Enteral omega-3 fatty acid, gamma-linolenic acid, and antioxidant supplementation in acute lung injury. *JAMA,* 306, 1574-1581.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: with application in R*.

Roozenbeek, B., Lingsma, H.F., Steyerberg, E.W., Maas, A.I., & Group, I.S. (2010). Underpowered trials in critical care medicine: how to deal with them? *Crit Care,* 14, 423.

Roozenbeek, B., Maas, A.I., Lingsma, H.F., Butcher, I., Lu, J., Marmarou, A., et al. (2009). Baseline characteristics and statistical power in randomized controlled trials: selection, prognostic targeting, or covariate adjustment? *Crit Care Med,* 37, 2683-2690.

Rosenbaum, P.R. (2006). Comment: The place of death in the quality of life. *Statistical Science,* 21, 313-316.

Rubenfeld, G.D. (2015). Confronting the frustrations of negative clinical trials in acute respiratory distress syndrome. *Ann Am Thorac Soc,* 12 Suppl 1, S58-63.

Rubenfeld, G.D., & Abraham, E. (2008). When is a negative phase II trial truly negative? *Am J Respir Crit Care Med,* 178, 554-555.

Rubin, E.B., Buehler, A.E., & Halpern, S.D. (2016). States Worse Than Death Among Hospitalized Patients With Serious Illnesses. *JAMA Intern Med*.

Sackett, D.L., Richardson, W.S., Rosenberg, W., & Haynes, R.B. (1997). *Evidence-based medicine: How to practice and teach EBM.* New York: Churchill Livingston.

Scales, D.C. (2013). Research to inform the consent-to-research process. *Intensive Care Med,* 39, 1484-1486.

Scales, D.C., & Rubenfeld, G.D. (2005). Estimating sample size in critical care clinical trials. *Journal of critical care,* 20, 6-11.

Schoenfeld, D.A., Bernard, G.R., & Network, A. (2002). Statistical evaluation of ventilator-free days as an efficacy measure in clinical trials of treatments for acute respiratory distress syndrome. *Crit Care Med,* 30, 1772-1777.

Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., et al. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA,* 315, 801-810.

Sjoding, M.W., Luo, K., Miller, M.A., & Iwashyna, T.J. (2015). When do confounding by indication and inadequate risk adjustment bias critical care studies? A simulation study. *Crit Care,* 19, 195.

Spragg, R.G., Bernard, G.R., Checkley, W., Curtis, J.R., Gajic, O., Guyatt, G., et al. (2010). Beyond mortality: future clinical research in acute lung injury. *Am J Respir Crit Care Med,* 181, 1121-1127.

Stevenson, E.K., Rubenstein, A.R., Radin, G.T., Wiener, R.S., & Walkey, A.J. (2014). Two decades of mortality trends among patients with severe sepsis: a comparative meta-analysis*. *Crit Care Med,* 42, 625-631.

Tomlinson, G., & Detsky, A.S. (2010). Composite end points in randomized trials: there is no free lunch. *JAMA : the journal of the American Medical Association,* 303, 267-268.

Tovey, D. (2011). The impact of Cochrane Reviews. *Cochrane Database Syst Rev,* 2011, ED000007.

Tritapepe, L., De Santis, V., Vitale, D., Guarracino, F., Pellegrini, F., Pietropaoli, P., et al. (2009). Levosimendan pre-treatment improves outcomes in patients undergoing coronary artery bypass graft surgery. *Br J Anaesth,* 102, 198-204.

Van den Berghe, G., Wilmer, A., Hermans, G., Meersseman, W., Wouters, P.J., Milants, I., et al. (2006). Intensive insulin therapy in the medical ICU. *N Engl J Med,* 354, 449-461.

van Meurs, M., Ligtenberg, J.J., & Zijlstra, J.G. (2008). The randomized controlled trial needs critical care. *Crit Care Med,* 36, 3118-3119; author reply 3119.

Villanueva, C., Colomo, A., Bosch, A., Concepcion, M., Hernandez-Gea, V., Aracil, C., et al. (2013). Transfusion Strategies for Acute Upper Gastrointestinal Bleeding. *New England Journal of Medicine,* 368, 11-21.

Vincent, J.-L. (2010). We should abandon randomized controlled trials in the intensive care unit. *Critical Care Medicine,* 38, S534-S538.

Vincent, J.L. (2009). Logistics of large international trials: the good, the bad, and the ugly. *Crit Care Med,* 37, S75-79.

Wagner, J., Gabler, N.B., Ratcliffe, S.J., Brown, S.E., Strom, B.L., & Halpern, S.D. (2013). Outcomes among patients discharged from busy intensive care units. *Ann Intern Med,* 159, 447-455.

Wendler, D., & Rid, A. (2011). Systematic Review: The Effect on Surrogates of Making Treatment Decisions for Others. *Annals of Internal Medicine,* 154, 336-346.

Williamson, P., & Clarke, M. (2012). The COMET (Core Outcome Measures in Effectiveness Trials) Initiative: Its Role in Improving Cochrane Reviews. *Cochrane Database Syst Rev,* 5, ED000041.

Yang, F., & Small, D.S. (2016). Using post-quality of life measurement information in censoring by death problems. *Journal of the Royal Soceity of Statistics: Series B*.

Young, P., Hodgson, C., Dulhunty, J., Saxena, M., Bailey, M., Bellomo, R., et al. (2012). End points for phase II trials in intensive care: recommendations from the Australian and New Zealand Clinical Trials Group consensus panel meeting. *Crit Care Resusc,* 14, 211-215.

Zimmerman, J.E., Kramer, A.A., McNair, D.S., Malila, F.M., & Shaffer, V.L. (2006). Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. *Crit Care Med,* 34, 2517-2529.