Empirical Bayes Methods for Modern Statistical Problems

By

Derek K Smith

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biostatistics

May, 2017

Nashville, Tennessee

Approved:

Professor William Dupont , Ph.D.

Professor Jeffrey Blume , Ph.D.

Professor Robert Greevy , Ph.D.

Professor Sonya Sterba , Ph.D.

This work and the progress I have made over the last several years would not have been possible without the constant support of my family Loren, Arabel, and Isaiah who have all made sacrifices so that I could pursue this work. In addition, I am grateful for the support of my parents Glenn and Donna Becker without whom returning to student life while becoming a parent myself would have been all the more difficult.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

This work develops an empirical Bayes approach to statistical difficulties that arise in real-world applications. Misuse of these methods as though the resulting posterior distributions were true Bayes posteriors has lead to limited adoption. The first problem solved via an empirical Bayes approach deals with surrogate outcome measures In this work we propose criteria similar to the Prentice criteria for using surrogates to develop risk scores. Their behavior is investigated through a series of simulation studies and an empirical Bayes weighting scheme is developed which alleviates their pathologic behavior. It is then hypothesized that a common clinical measure, change in perioperative serum creatinine level from baseline, is actually a partial surrogate. The result is a more acurate predictive model for both short and long-term measures of kidney function. The second problem solved deals with likelihood support intervals. Likelihood intervals are a way to quantify statistical uncertainty. In this work we develop a novel procedure based on the bootstrap for estimating the frequency characteristics of likelihood intervals. The resulting intervals have both the frequency properties of the set as well as each individual member of the set attaining a specified support level. An R package, *supportInt*, was developed to calculate these intervals and published on the Comprehensive R Archive Network. The third problem addressed deals with the design of clinical trials when the potential protocols for the intervention are highly variable. It is demonstrated that large single protocol designs that are frequently advocated for can be replaced by multi-arm protocols to more accurately assess the question of an interventions potential efficacy. Simulation studies are conducted that make use of a novel adaptive randomization scheme based on an empirically estimated likelihood function. A Shiny app allows for the conduct of further studies by the reader under a wide variety of conditions.

Chapter 1

Introduction

The methods presented in this dissertation revolve around empirical data methods. In each situation we are faced with a unique problem that can be addressed by using an empirical Bayes approach. The empirical Bayes approach is then compared to more traditional methods to determine what advantages or disadvantages it may demonstrate. Ultimately in each case the empirical Bayes approach demonstrates a potential for large gains, however, these gains are sometimes dependent on the particular situation.

1.1 Introduction to Empirical Bayes Methods

1.1.1 Robbins and Early Empirical Bayes Methods

The orgins of empirical Bayes estimation begin in the mid 1950s, with Herbert Robbins pioneering the area of empirical Bayes point estimation. In 1956 he outlined how empirical Bayes estimates would also be expected to dominate the usual MLE (1). His formulation of the problem begins with a discrete random variable $X \sim p(X = x|\Lambda = \lambda)$ dependent on a parameter Λ, which is itself randomly distributed $\Lambda \sim G(\lambda) = P(\Lambda \leq \lambda)$. He is concerned with finding the realization of Λ, call it λ, that minimizes the expected value of a specified loss function. The expected value of a loss function is called risk. In this case the loss function Robbins seeks to minimize is squared-error loss. Then he shows that for any estimate of the random quantity Λ, say $\varphi(x)$,

$$\begin{aligned} E_X[\varphi(X) - \Lambda]^2 &= E_\Lambda[E_{X|\Lambda}[(\varphi(X) - \Lambda)^2]] \\ &= \int \sum_x p(x|\lambda)[\varphi(x) - \lambda]^2 dG(\lambda) \end{aligned}$$

is minimized by,

$$\varphi(x) = \frac{\int \lambda p(x|\lambda)dG(\lambda)}{\int p(x|\lambda)dG(\lambda)} = \int \lambda p(\lambda|x) = E[\Lambda|X]$$

In other words it is the Bayesian posterior mean that minimizes squared error loss.

Proof: The expected squared error loss is minimized by:

$$\begin{gathered} argmin_{\varphi(x)} \int p(x|\lambda)[\varphi(x)-\lambda]^2 dG(\lambda) \\ = argmin_{\varphi(x)} \int p(x|\lambda)dG(\lambda)\left(\varphi(x) - \tfrac{\int \lambda p(x|\lambda)dG(\lambda)}{\int p(x|\lambda)dG(\lambda)}\right)^2 + \left(\int \lambda^2 p(x|\lambda)dG(\lambda) - \tfrac{(\int \lambda p(x|\lambda)dG(\lambda))^2}{\int p(x|\lambda)dG(\lambda)}\right) \\ = argmin_{\varphi(x)} \int p(x|\lambda)dG(\lambda)\left(\varphi(x) - \tfrac{\int \lambda p(x|\lambda)dG(\lambda)}{\int p(x|\lambda)dG(\lambda)}\right)^2 + c \\ \Rightarrow \varphi(x) = \tfrac{\int \lambda p(x|\lambda)dG(\lambda)}{\int p(x|\lambda)dG(\lambda)} = E[\Lambda|X] \end{gathered}$$

Having illustrated that the posterior mean is optimal in mean square error (MSE), he then addresses the practicality that the prior of $\Lambda, G(\lambda)$, is largely unknown in real-world problems making the posterior mean difficult to estimate in many applied scenarios. He suggests that if there were a sequence of experiments available,

$$(X_1,\Lambda_1),(X_2,\Lambda_2),\ldots,(X_n,\Lambda_n)$$

that one could substitute the marginal empirical distribution (ecdf) of the combined $X = (X_1,\ldots,X_n)$ for the actual marginal distribution $\int p(x|\lambda)dG(\lambda)$. This is reasonable because the ecdf $\hat{F}_n(x) \overset{a.s.}{\to} p(x)$, by the strong law of large numbers. Robbins then shows how to calculate empirical Bayes estimates and demonstrates some superiority in MSE for the Poisson, Geometric, Binomial, and Laplace distribution families.

1.1.2 The James-Stein Estimator

In 1955 Charles Stein shocked the statistical community. He showed that when estimating 3 or more means, from normally distributed populations, the vector of sample means is actually inadmissible with respect to squared error loss (2). An estimator $\hat{\theta}$ is inadmissible if there is another estimator $\widetilde{\theta}$ that has lower expected squared error loss (i.e., risk) for every value of θ. At the time, the vector of maximum likelihood estimators was widely believed to be optimal; after all, one or two sample means was known to be admissible under squared error loss, so why not 3 or more? This counter-intuitive result came to be known as Steins Paradox.

James and Stein (3) provided an estimator that dominated the vector of sample means, but their suggestion has since been shown to be inadmissible as well. Some years later, Efron and Morris (1973) recognized that, under certain fairly general conditions, Stein-type estimators were equivalent to employing an Empirical Bayes estimation scheme (4). An empirical Bayes estimator is an estimate derived from the posterior distribution of a Bayesian model where the hyperparameters, the parameters that define the prior distribution, are estimated from the data. This is not a traditional Bayes approach, which requires that the statistician take into account their uncertainty about the hyperparameter. Moreover, the empirical prior is envisioned as a tangible marginal distribution that can, in theory, be observed. Empirical Bayes procedures are also considered suspect for "using the data twice": Once to estimate the prior and again in the likelihood. Although much work has been done to develop valid confidence methods for empirical Bayes estimators, the highly technical nature of many of the proposed solutions has resulted in limited utilization in many areas of research.

Efron-Morris

In addition to identifying the empirical Bayes structure of James-Stein estimator, Efron and Morris (1973) also proposed a slight modification to the procedure in which MLEs would be shrunk toward the grand mean as opposed to zero (4). Thus the Efron-Morris estimate is given by the posterior mean of the model,

$$x_{ij}|\mu_i \sim N(\mu_i, \sigma^2) \quad \mu_i \sim N(\bar{\bar{x}}, \tau^2)$$

and is given by

$$\hat{\theta}_i^{EM} = \bar{\bar{X}} + \left(1 - \frac{k-3}{\sum(\overline{X_i} - \bar{\bar{X}})^2}\right)(\overline{X_i} - \bar{\bar{X}})$$

In 1975 Efron and Morris published an applied example using their proposed estimate (5). This example involving baseball batting averages was one of the first and highest impact tutorials in practical usage of empirical Bayes estimates.

Random Effects Regression

Stein's phenomenon also applies in the regression setting. The empirical Bayes predictions from a random effects model are the best linear predictors of the individual mean values (6). As mentioned in Laird and Ware 1982, the model

$$\boldsymbol{y_i} = \boldsymbol{X_i \alpha} + \boldsymbol{Z_i b_i} + \boldsymbol{e_i}$$

where $\boldsymbol{e_i} \sim N(\boldsymbol{0}, \boldsymbol{R}_i)$ and $\boldsymbol{b_i} \sim N(\boldsymbol{0}, \boldsymbol{D})$ gives rise to the empirical Bayes prediction

$$\hat{\boldsymbol{b_i}} = \boldsymbol{D Z_i^t W}_i(\boldsymbol{y_i} - \boldsymbol{X_i \hat{\alpha}})$$

Where $\boldsymbol{W}$ is the inverse of the covariance of the y's (7). In this case $\hat{\boldsymbol{b_i}} = E[\boldsymbol{b}_i|\boldsymbol{y_i}, \boldsymbol{\hat{\alpha}}, \boldsymbol{\theta}]$, where $\boldsymbol{\theta}$ is a q-vector of variance covariance parameters. Because the prior mean of the random-effects distribution is commonly assumed to be zero, the resulting $\hat{b_i}$ is an average of the ordinary least squares fixed effect and the zero vector (because the residuals have mean zero by design). This makes the estimates quite similar to James-Stein type estimators.

The Bootstrap as an Empirical Bayes Process

Point estimation is not the only area where empirical Bayes methods can be used to improve one's understanding of an applied statistical problem. The bootstrap has become an extremely popular method for analyzing complicated data. The basic idea behind the bootstrap, commonly referred to as the "plug-in principle" or the "bootstrap principle", begins with a goal of estimating some functional of the true unknown distribution, $\theta(F_0)$. Since the true distribution is unknown, a sample $(x_1,\ldots,x_n)$ is collected from the population and F_0 can be estimates and used to calculate $\theta(\hat{F})$. $\hat{F}$ is typically taken to be the empirical distribution function resulting in the non-parametric maximum likelihood estimate (8). For ease of notation $\hat{F}$ and F_1 will be used interchangably. The bootstrap principle suggests plugging in F_1 will likely result in good estimates, depending on the resemblance between F_0 and F_1 (9).

Although this initial point estimation is the simplest instance of the plug-in principle, it is far from the most useful. Now suppose that there is a bias associated with our estimate, $\theta(F_1)$, which requires an adjustment. Instead of using $\theta(F_1)$ as our estimate of $\theta(F_0)$, we could introduce an additive correction, t, giving

$$\theta(F_0) \approx \theta(F_1) + t$$

where

$$t = \theta(F_0) - E[\theta(F_1)|F_0].$$

Although the ability to make this correction would be ideal in practice, the definition of

t again requires that the analyst know the true value of $\theta(F_0)$. The bootstrap principle suggests a potential solution replacing F_i with F_{i+1}. This eliminates the need for knowledge of F_1 but requires the estimation of $E[\theta(F_2)|F_1]$, which is done using a bootstrap resampling scheme. The sample data is resampled with replacement B times. Each time a new $\theta(F_{2,i})$ is calculated. When the sampling is complete the estimate is then calculated as

$$\hat{E}[\theta(F_2)|F_1] = \frac{1}{B}\sum_{i=1}^{B}\theta(F_{2,i}).$$

We can now estimate t using the bootstrap principle as

$$\hat{t} = \theta(F_1) - E[\theta(F_2)|F_1]$$

and alter our point estimate accordingly.

Understanding the bootstrap/plug-in principle is the key to understanding the bootstrap as an empirical Bayes procedure. This principle not only guides the application of the bootstrap, but was the central idea in Robbins' original presentation of empirical Bayes methods (recall that his formulation called for writing the posterior as a function of the unknown prior predictive distribution and then replacing it with the empirical distribution). A deeper understanding of this relationship requires a slightly different formulation of the bootstrap as repeated multinomial draws from the observed data.

Suppose that as the analyst I want to conduct a Bayesian analysis of my data, ultimately calculating my estimate of $\theta(F_0)$ from the posterior distribution. Also suppose there is a discrete distribution F_X that places a certain probability p_i on each point in the support. A logical model to consider would be the Dirchlet-multinomial conjugate model, which gives

$$p \sim Dir(\alpha)$$

$$f(x|p) = \prod_{i=1}^{n} p_i, i|x_i \in x_1, \ldots, x_n$$

$$p|x \sim Dir(\alpha + N_{\{i|x_i \in x_1,\ldots,x_n\}})$$

where $N_{\{i|x_i \in x_1,\ldots,x_n\}}$ is a vector corresponding to the number of times a particular point in the support of X appeared in the sample. Now consider a bootstrap analysis of the same data which would yield

$$x^* \sim \frac{1}{n} Mult(n, \hat{p})$$

since the bootstrap samples are multinomial draws on the observed data. This is very similar to the result of the Bayes analysis when $\alpha \approx 0$, i.e. when the prior weight is evenly distributed over the observed data points. Thus bootstrap methods can be thought of as approximating samples from the posterior predictive result from an empirically defined prior distribution.

1.1.3 Latent Class Models as Empirical Bayes Methods

Another application in modern statistics that relies heavily on empirical Bayes methods is the latent class model. Latent class models are useful when there is a categorical covariate that is unmeasured, but crucial for the model's success. These models rely heavily on the expectation maximization (EM) algorithm to obtain maximum likelihood estimates. As an example consider a binary latent class model. The EM fitting process starts with random class assignments and updates the probability of class membership using the fitted likelihood for two classes. In other words there is an estimated prior probability of class assignment that is iteratively updated along with the likelihod fits as the EM algorithm progresses. When convergence is achieved, the result is estimates of the prior probability of class assignment which are updated via the likelihoods to yield an empirical Bayes posterior estimate of class membership.

Chapter 2

Modeling Partial Surrogate Outcomes: applications to acute kidney injury

2.1 Introduction to the Chapter

This chapter begins by addressing an applied problem, and expands on the statistcal ideas that must be considered in order to properly address the problem. I begin with an explanation of the applied scenario which served as the inspiration for this section. This example will serve as motivation for a general statistical issue that has gone unrecognized in the clinical literature to this point. Having elaborated on the statistical issues, simulated examples will be used to illustrate the statistical principles that allow the problem to be addressed. Finally, I will return to the original problem and attempt to demonstrate how this novel statistical approach improves the modeling of relavent clinical outcomes.

2.2 Acute Kidney Injury

Acute kidney injury (AKI) can refer to any number of situations in medicine. It is meant to describe a patient that has experienced a physiologic stress resulting in damage to the kidneys. This damage can refer to structural damage or a decreased filtration capacity. Historically the way this has been assessed clinically is using a functional biomarker called serum creatinine and the need for hemodialysis. AKI has been linked to a host of adverse clinical outcomes including increased length of stay, development or progression of chronic kidney disease (CKD), and mortality (10; 11).

Creatinine is a byproduct produced by the breakdown of muscle in the body. As a person goes about their daily business the production and removal of creatinine from the serum reaches a steady state which is representative of the kidney's filtration. This measurement at steady state provides the definition for chronic kidney disease by providing an estimate of a patient's glomerular filtration rate at baseline. In an acute setting however, it is unclear how good of a marker serum creatinine is when it comes to assessing the condition of the kidney. When exposed to physiological stress that causes kidney damage, the kidneys have the ability to transiently increase their functionality which masks the injury from any change in a functional biomarker, such as serum creatinine. This ability to increase filtration is commonly known as functional renal reserve. Perhaps the most poignant example of serum creatinine's failure to detect nephron damage is that it is not uncommon for patients who are donating a kidney for transplant to not experience any increase in their serum creatinine despite having permanently lost roughly 50% of their nephrons as a result of the surgical procedure.

Despite its apparent limitations as a marker of injury in the perioperative period, serum creatinine elevation has been the centerpiece of virtually every attempt to define or detect AKI. The most widely known criteria (RIFLE, AKIN, and KDIGO) by which AKI is defined all contain elevated serum creatinine to some degree as their most sensitive marker of damage (12; 13; 14). Attempts in the clinical literature to either identify clinical factors that are associated with nephron damage or to predict which patients will develop it have largely utilized multivariate logistic regression models whose outcomes are one of these three most popular criteria, which largely represent a dichotomization of serum creatinine change with minor contributions of extreme damage markers like the development of new dialysis. Due to the degree to which AKI has become synonymous with serum creatinine elevation in clinical circles and for the sake of clarity, for the remainder of this chapter "AKI" will refer to experiencing an acute serum creatinine increase and "nephron damage" will refer to sustaining some physical kidney damage that may or may not result in serum creatinine increase.

The primary question of clinical interest is whether it is possible to overcome serum creatinine's deficiency as a biomarker in order to makes legitimate inferences about likely associations with nephron damage and how best to predict it. For the purposes of this chapter we will be focused on perioperative AKI and nephron damage. Perioperative AKI is when the damage occurs during the time directly after a surgical intervention and is often attributed to the stress of the surgery.

2.3 Statistical Issues with Modeling AKI

The first and most obvious statistical issue with the modeling of AKI to this point is the shear amount of information that is discarded through the arbitrary dichotomization of serum creatinine change. This approach inappropriately treats minor and severe cases of AKI as though they were the same. For the duration of the chapter examples will revolve around modeling AKI continuously using linear models.

In the previous section, I have given a very brief summary of the biological considerations of AKI. In statistical terms this biology suggests that serum creatinine change is a surrogate measure of nephron damage, which is not easily observable. As a surrogate there are certain criteria that must be upheld in order for serum creatinine change as an outcome to be a valid surrogate for nephron damage. In 1989, Prentice described a set of four criteria sufficient to ensure valid hypothesis tests (15).

1. The proposed risk factor must be related to the surrogate. $f(S|Z) \neq f(S)$
2. The proposed risk factor must be related to the true outcome. $f(T|Z) \neq f(T)$
3. The surrogate must be related to the true outcome. $f(T|S) \neq f(T)$
4. The risk factor must be related to the true outcome only through the surrogate. $f(T|S,Z) = f(T|S)$

Although these criteria, when satisfied, assure legitimate hypothesis testing they are often difficult or impossible to verify in practice and therefore often have to be taken on assumption based on the best available scientific knowledge. This is especially true in cases where the true outcome is impossible to measure such as with nephron damage. Reviewing the criteria with serum creatinine change in mind as a surrogate of nephron damage, suggests that with careful covariate selection Prentice criteria 1, 2, and 3 are quite likely satisfied. Criterion 4 on the other hand is almost certainly in gross violation. Due to the previously described functional renal reserve, we expect that there are patients experiencing real nephron damage which do not manifest changes in the surrogate, serum creatinine. This would mean that the covariates Z would contain information about the truth that was not contained in S, violating the 4th criterion. This violation means that hypothesis testing with serum creatinine change as an outcome is in no way guaranteed to have the nominal error rate as a test for associations between covariates and nephron damage.

To reiterate, criterion 4 failed due to the fact that nephron injuries can be masked from serum creatinine changes when the amount of damage fails to exhaust the patient's functional renal reserve. However, if we were to restrict our population under study to those whose functional renal reserve was exhausted criterion 4 would be quite plausible. In other words, there exists a subpopulation in which serum creatinine change is likely a valid surrogate for nephron damage. In the complement population the effects of nephron damage on serum creatinine are masked producing a violation of not only Prentice's 4th criterion but likely criteria 1 and 3 as well.

This concept that a measure could be a valid surrogate in one subpopulation and an invalid surrogate in a complementary subpopulation has lead me to adopt the term *partial surrogate*.

2.4 Partial Surrogates

Definition: A **partial surrogate** is a measure that satisifies Prentice's four criteria for a given subpopulation but does not in its complementary subpopulation.

The statistical approach to dealing with a partial surrogate is situationally dependent. For example, suppose you had an indicator variable in your data set that defines the valid and invalid surrogate subpopulations. Then an analyst can simply conduct the analysis in the valid surrogate subpopulation by performing the appropriate regression:

$$Y_i = (1 - I_{invalid})(X_i\beta + \varepsilon_i) + I_{invalid}(X_i\gamma + \delta_i), \;\; \varepsilon \sim N(0,\sigma^2), \; \delta \sim N(0,\tau^2).$$

If the subpopulation information is unavailable as in the AKI example however, a more sophisticated approach is required. I propose the use of a latent variable regression model to attempt to distinguish the subpopulations and achieve valid inference. In 1977 Dempster, Laird, and Rubin demonstrated how the expectation-maximization (EM) algorithm could be used to obtain maximum likelihood estimates when data is missing under specific conditions (16).

In either case the analyst must be explicit about what population they believe the results are generalizable to. The population consideration is due to the fact that there is no guarantee that the valid subpopulation is a representative sample of the entire population. Other systematic differences between the populations effect to what population the result can be generalized to. Considering serum creatinine change as an example, we speculate that the valid surrogate subpopulation consists of the patients whose nephron damage exceeds their functional renal reserve. It is therefore easy to imagine this subpopulation representing those with some combination of severe nephron damage and compromised baseline renal function. Therefore after fitting the model it is necessary to consider whether science suggests extrapolation to other populations is reasonable or whether covariate realtionships with nephron damage may differ in lower risk populations.

2.4.1 Ignoring the Partial Surrogate

In order to motivate the discussion of the treatment of partial surrogates in the nontrivial case where the subgroup indicator variable is missing, consider the consequence of ignoring the partial nature of the surrogate as it has been done in the clinical literature on perioperative AKI to this point. Again we will restrict consideration to the linear model of the continuously valued outcome. Suppose that there is a single parameter of interest β whose relationship with the true outcome, T, is given by

$$t_i = \beta x_i + \varepsilon_i, \;\; \varepsilon_i \sim N(0,1).$$

Now assume that for this particular problem T cannot be observed. Instead, S is observed which is related to T. To draw further analogy to the AKI problem, suppose that S is equal to T in a subpopulation and in the complementary population of proportion p, S is a standard normal deviate. If we ignore the partial surrogate nature of S, we would fit the linear model $E[S] = X\beta$ and the maximum likelihood estimate would have expectation

$$\begin{gathered} E[(X'X)^{-1}X'S] = (X'X)^{-1}X'E[S] \\ = (1-p)(X'X)^{-1}X'X\beta + p(X'X)^{-1}X'E[Z], \;\; Z \sim N(0,1) \\ = (1-p)\beta + 0 = (1-p)\beta. \end{gathered}$$

Ignoring the partial nature of the surrogate outcome results in a biased estimate of the parameter of interest. In cases where the subpopulations are defined by a threshold effect below which the surrogate is expected to produce a null value, this bias is manifested as a null-bias. In the case of linear models the degree of bias is given by the size of the null subpopulation. In the case of perioperative AKI (AKI in the time surrounding surgery) this null subpopulation accounts for around $70-80\%$ of the entire population. This large null-bias effect combined with the loss of power resulting from the unnecessary dichotomization of serum creatinine change explains why attempts in to identify relevant associations between supposed risk factors and AKI in clinical research have yielded inconsistent results even in studies that appear to be well-powered with respect to serum creatinine change.

To illustrate this point further, consider again the above scenario in which the true data generating model is a linear relation. A simple simulation can demonstrate both the loss in power and the bias introduced by the partial surrogate as a function of the residual variance and subpopulation proportion, p, respectively. In the simulation 1000 data points are generated from the true linear model. In the first simulation the residual variance, the variance of ε, is allowed to vary from 5^2 to 15^2 while the subpopulation proportion is fixed at 75% of observations being reduced to standard normal deviates. In the second simulation, the residual variance is fixed at 10 and the subpopution proportion, p, is allowed to vary from 10% to 90%. In both simulations 5000 replicates are performed for each experimental situation.

These simulations clearly demonstrate the price that is paid for ignoring the partial surrogate nature of, S. Depending on the structure of the problem, i.e. the $Var(\varepsilon)$ and p, the analyst will be operating at substantially reduced power and be obtaining biased estimates of β even in this case where in the valid subpopulation $S = T$, which will not necessarily be true generally, see Figure 2.1.

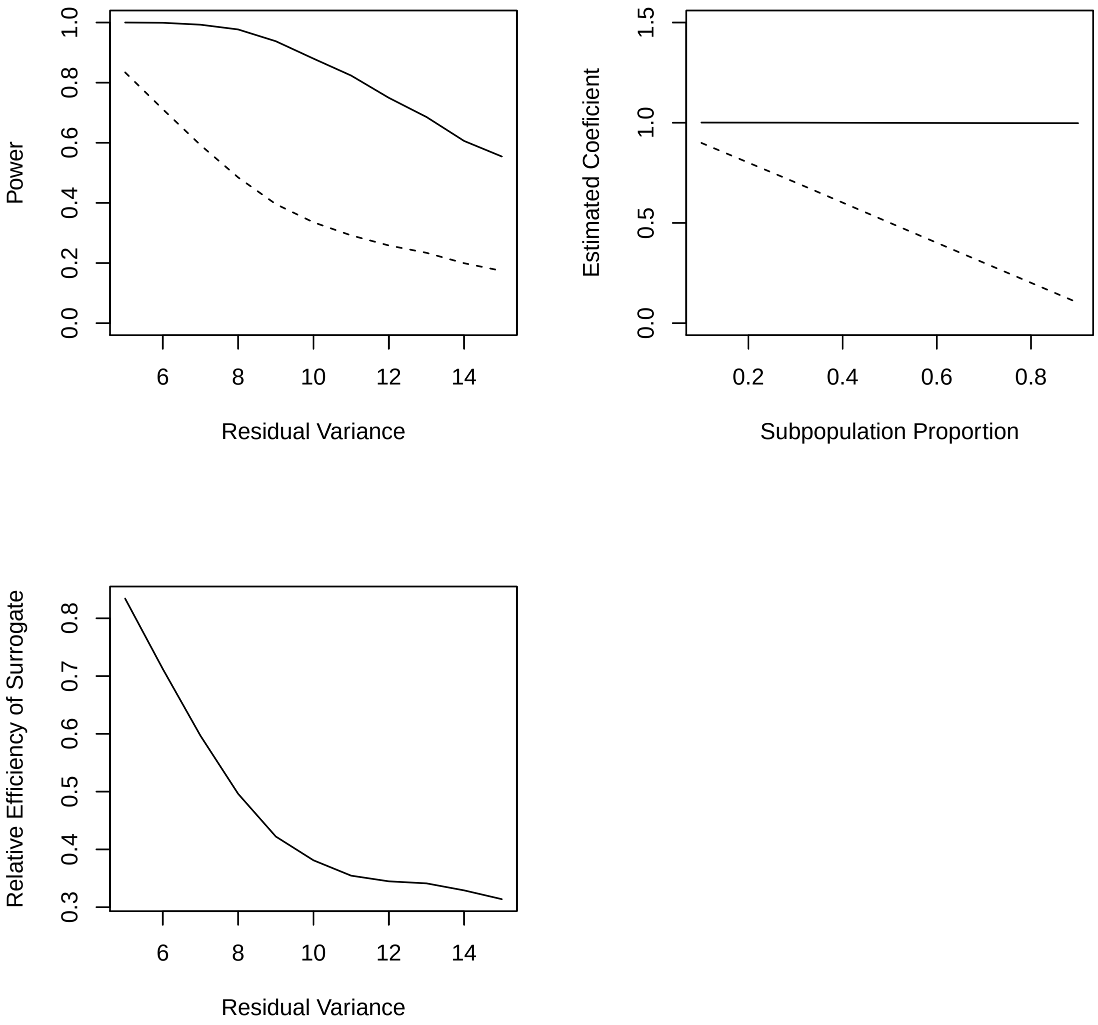


Figure 2.1: UL: Difference in power between the analysis of the true outcome (solid) and the analysis of the surrogate outcome (dashed) over various values of residual variance. LL: Relative efficiency of surrogate analysis to the true outcome analysis over various values of residual variance. UR: Difference in expected value of the coeficient estimate over various subpopulation proportions, *p*, between the true outcome analysis (solid) and the surrogate outcome analysis (dashed).

2.4.2 Latent Variable (Mixture) Models for Partial Surrogates

The mixture model approach to dealing with a latent categorical variable, in the case of AKI this is subpopulation ID, is well developed and will only be briefly summarized here. As previously stated, if the subpopulation were known it would be a simple matter to conduct the analysis within the subpopulation for which the surrogate is valid, or to perform a stratified analysis using the model

$$Y_i = (1 - I_{invalid})(X_i\beta + \varepsilon_i) + I_{invalid}(X_i\gamma + \delta_i), \;\; \varepsilon \sim N(0, \sigma^2), \; \delta \sim N(0, \tau^2).$$

Given that certain conditions hold, The mixture model approach provides maximum likelihood estimates for β despite the missingness of the subpopulation indicator that precludes the fitting of the above model. This is accomplished by using the expectation maximization algorithm (16). The EM algorithm is an iterative algorithm that sequentially calculates the expected log-likelihood given some current parameter values and then calculates new parameter values by maximizing the expectation over the parameters. This process is continued until the algorithm converges to a maximum with care taken to ensure that the attained maximum is a global one.

In the case of a partial surrogate, the latent factor we are concerned with defines the subpopulation in which S is a valid surrogate for T. The goal of the EM procedure is to come up with a probability that a given patient is from one subpopulation or the other. In clinical problems these subpopulations will define two phenotypes exhibited by patients. For example, in the AKI problem we expect that one of the two phenotypes will exhibit little to no relationship between preoperative and intraoperative patient characteristics and the surrogate, serum creatinine elevation, whereas the other will suggest clinically meaningful relationships. The more distinct and less variable these phenotypes are, the more reliable the EM procedure will be at assigning probabilities that a given patient is of a particular phenotype.

The mixture model that results from the EM-procedure is an empirical Bayes type model. It consists of two linear regression models representing the two phenotypes described above. These two models are combined by the subpopulation proportion p.

$$Y_i \sim \hat{p} f_1(Y|X) + (1-\hat{p}) f_2(Y|X)$$

This model weights the two linear models by the estimated subpopulation proportion. This estimate of the subpopulation proportion is know as the prior mixing proportion and is updated at each step of the EM algorithm. This prior proportion is then combined with with the likelihood ratio of the two linear model pieces, the relative probability that a patient exhibits a particular phenotype, to give a patient specific probability of belonging to a given phenotype, which I shall denote $\hat{p}_i$.

$$\hat{p}_i = \frac{o_i}{1+o_i}, \;\; o_i = \frac{1}{1 + \frac{\hat{p}}{1-\hat{p}} \frac{f_1(Y_i|X_i)}{f_2(Y_i|X_i)}}$$

In the AKI example $\hat{p}$ and $\hat{p}_i$ are defined to represent the marginal (prior) and person specific (posterior) probabilities of belonging to the phenotype that shows little to no relation between suspected risk factors and serum creatinine change.

These person specific probabilities are used to weight each participant's contribution to each of the phenotype linear models. If patient $\hat{p}_4 = 0.80$ on the last iteration of the EM algorithm then patient 4 will contribute 80% weight to the model estimating the null phenotype and 20% weight to the model estimating the valid surrogate phenotype. The phenotype models can thus be estimated via weighted least squares as

$$min_\beta \left(\sum_{i=1}^{n} (1-\hat{p}_i)(y_i - x_i\beta)^2 \right) \;\; and \;\; min_\gamma \left(\sum_{i=1}^{n} \hat{p}_i (y_i - x_i\gamma)^2 \right).$$

Fitted values can be obtained for the mixture by weighting the predicted values from the two phenotype models by their personal empirical Bayes posterior probabilities of being from the appropriate phenotype.

When the phenotypes defined by the partial surrogate subpopulations are sufficiently distinct and precise, it is possible to obtain maximum likelihood estimates and corresponding hypothesis tests of suspected risk factors within the phenotype defined by the valid surrogate subpopulation regardless of whether the subpopulation ID variable is available. The analyst must then utilize scientific knowledge to determine which populations the result is generalizable to.

2.5 Simulated Examples

2.5.1 Example 1

Simulations done in R version 3.2.0.

Having outlined the statistical principles behind using mixture models for making valid statistical inferences when presented with a partial surrogate, these principles will now be demonstrated via a simulated example. This example was designed to mimic the suspected biology in the AKI example. It contains two suspected risk factors, A and B. A is continuously values whereas B is binary. These risk factors are linearly related to the true outcome, T. Each patient is assigned a positively valued c_i representing their person specific functional renal reserve. The partial surrogate is equal to $T_i - c_i$ in 25% of the population and reduced to a Normal deviate in the remainder of the population.

$$A_i \sim 0.7N(0,1) + 0.3N(0.2,1)$$

$$B_i \sim Bern(0.7)$$

$$T_i = 0.518 + 0.3A_i + 0.8B_i$$

$$c_i \sim Gam(2,2)/5$$

$$U_i \sim Unif(0,1)$$

$$S_i = I_{[U_i>0.25]}N(0,0.3) + (1 - I_{[U_i>0.25]})(T_i - c_i)$$

500 samples were generated using this mechanism. The data were then analyzed using

linear regression of S on A and B and the corresponding mixture model approach.

	Estimate	Std. Error	Pr(>|t|)
Intercept	0.0509	0.0402	0.2062
A	0.0671	0.0285	0.0190
B	0.1989	0.0477	0.0000

Table 2.1: Results from linear model analysis of the surrogate, S.

As previously discussed ignoring the partial surrogate nature of S in this type of scenario results in a null bias in estimating the coeficients and overly conservative hypothesis testing, see Table 2.1. The BIC for the linear model was 719.55.

	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|t|)
Intercept	-0.0443	0.0231	0.0559	0.3585	0.0111	0.0000
A	0.0073	0.0161	0.6473	0.2988	0.0086	0.0000
B	0.0599	0.0274	0.0294	0.7999	0.0132	0.0000

Table 2.2: Results from mixture model analysis of the surrogate, S.

The same analysis was conducted with the mixture model approach. Because this example was designed with the biology of the AKI problem in mind and the surrogate is a nonuniform translation of the true outcome within the valid subpopulation (i.e. the surrogate is equal to $T_i - c_i$ when not reduced to nullity), it is expected that the mixture approach will result in one component with largely null associations and one component with unbiased estimates of the associations. Table 2.2 shows the results of the two mixture model components verifying this expectation. The BIC for the mixture model was 514.67. Calibration plots for both the linear and mixture model approach are given in Figure 2.2

This simulated example demonstrates how the mixture model can acheive accurate estimates of the data generating parameters by accounting for partial surrogate nature of S. It also shows the hazard of pooling estimates across phenotypes.

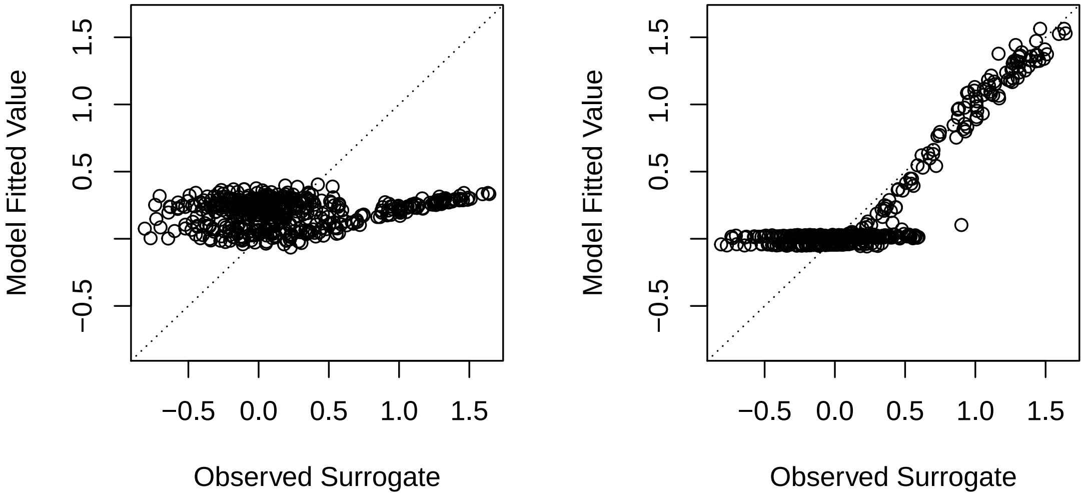


Figure 2.2: Calibration plots for the linear model fit (left) and the mixture model fit using posterior probabilities as weights (right)

2.5.2 Example 2

However, this method is not without its limitations. As previously noted, this method is dependent on the phenotypes being sufficiently different so that the EM procedure can determine a reliable probabilistic framework for which phenotype a given observation belongs to. In this second example the data generating mechanism is changed so that $T_i = 0.518 + .1A + .2B$ effectively reducing the distinction between the phenotypes. The analysis was then repeated as before.

	Estimate	Std. Error	Pr(>|t|)
Intercept	-0.0205	0.0217	0.3454
A	0.0307	0.0154	0.0466
B	0.0698	0.0258	0.0070

Table 2.3: Results from linear model analysis of the surrogate, S.

Again ignoring the partial surrogate nature of S in this type of scenario results in a null bias in estimating the coeficients and overly conservative hypothesis testing, see Table 2.3.

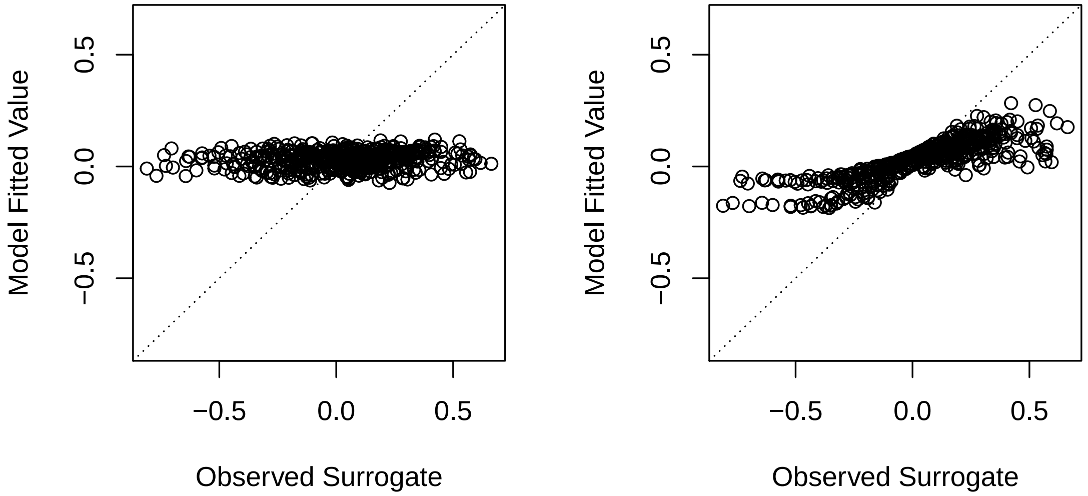


Figure 2.3: Calibration plots for the linear model fit (left) and the mixture model fit using posterior probabilities as weights (right)

The BIC for the linear model was 103.19.

	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|t|)
Intercept	0.1415	0.0146	0.0000	-0.1715	0.0217	0.0000
A	0.0718	0.0100	0.0000	-0.0120	0.0160	0.4540
B	0.0211	0.0172	0.2210	0.1074	0.0260	0.0000

Table 2.4: Results from mixture model analysis of the surrogate, S.

Table 2.4 shows the results of the two mixture model components. The BIC for the mixture model was 200.83. Calibration plots for both the linear and mixture model approach are given in Figure 2.2

In this case the EM procedure was unable to resolve the different phenotypes. Whereas in the previous example the BIC of the mixture model dominated that of the linear approach, in this example the mixture approach is substantially worse than the linear model due to the increased number of fit parameters. The failure of the EM procedure to distinguish the phenotypes is further evidenced by examining the distribution of posterior probabilities, $\hat{p}_i$, see Figure 2.4. The high density in the middle of the histogram represents a high degree of entropy, uncertainty of phenotype assignment, and can be used along with the BIC as a diagnostic for failure of the EM procedure.

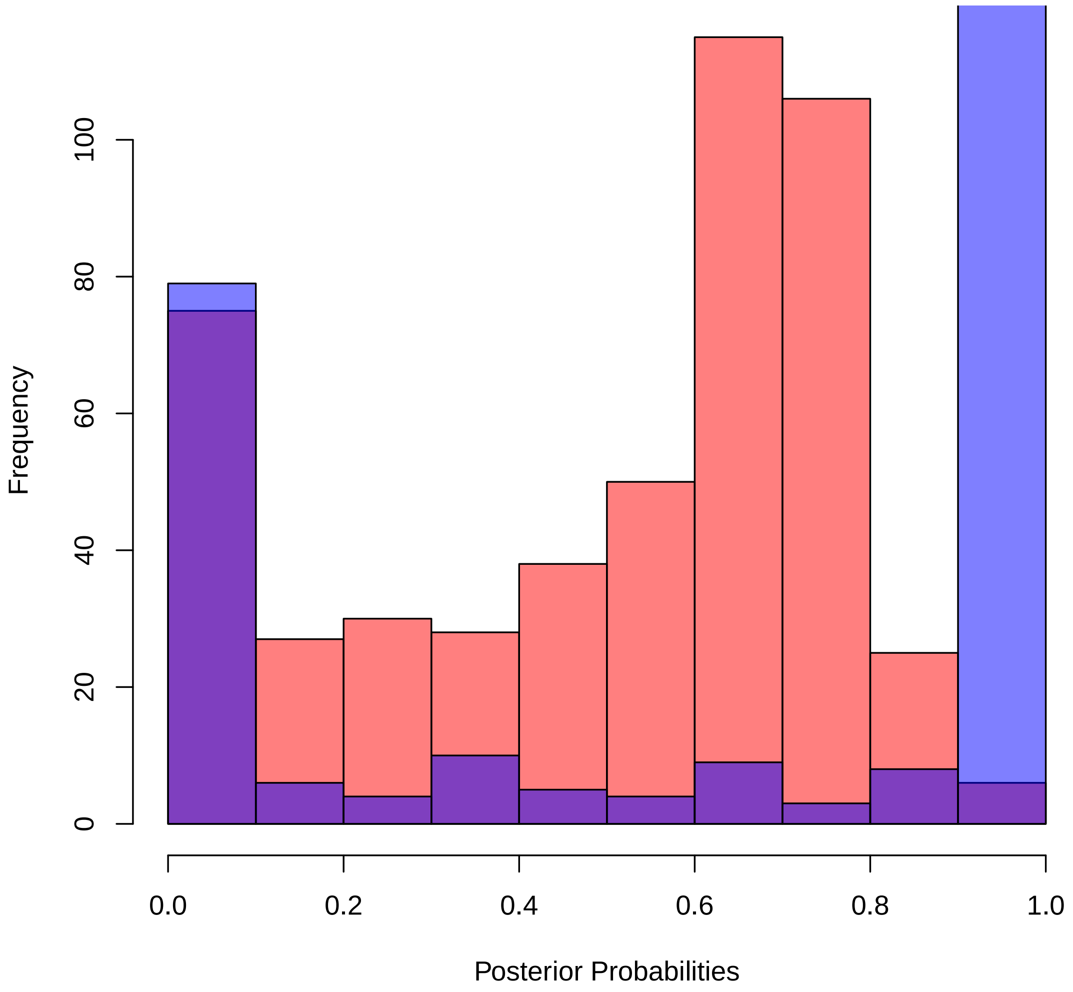


Figure 2.4: Distribution of posterior probabilities in the first (blue) and second (red) simulated examples.

2.6 Clinical Example

This clinical example is based on data collected as part of a clinical trial examining the effect of preoperative statin administration on the development of AKI following cardiac surgery. The dataset contains data from 541 patients. Suspected risk factors were identified by clinical subject matter experts. Serum creatinine measurements were made preoperatively and on the first postoperative day on all patients. Serum creatinine measurements were available on the second postoperative day on all but three patients. The primary outcome is the maximum serum creatinine change over the first two postoperative days. The data was analyzed via a linear and mixture model approach.

2.6.1 Linear Model Analysis of Perioperative AKI

The maximum serum creatinine change over the first two postoperative days was regressed on the suspected risk factors using a multivariate linear regression model, see Table 2.5. The BIC for the model was 525.38

	Estimate	Std. Error	Pr(>|t|)
Intercept	-0.3324	0.2890	0.2506
BMI	0.0100	0.0028	0.0005
Urine Output	-0.0001	0.0001	0.0250
Fluid Given	0.0000	0.0000	0.8304
Baseline SCr	0.2898	0.1710	0.0907
Age	0.0031	0.0033	0.3472
Length of Surgery	0.0006	0.0002	0.0091
Maximum Lactate Level	-0.0129	0.0130	0.3187
Cross Clamp Time	0.0007	0.0003	0.0491
Baseline - Mean Diastolic	-0.0013	0.0013	0.3298
Pulse Pressure	0.0011	0.0006	0.0742
Bypass Time	0.0003	0.0003	0.2558
Hespan	0.0002	0.0001	0.0337
Hypertension	0.0196	0.0362	0.5882
Diabetes	-0.0241	0.0364	0.5075
CHF	0.0009	0.0361	0.9801
COPD	-0.0697	0.0686	0.3106
HB	-0.0312	0.0092	0.0008
Baseline SCr*Age	-0.0018	0.0026	0.4860

Table 2.5: Results from linear model applied to clinical data.

2.6.2 Mixture Model Analysis of Perioperative AKI

The mixture model was applied to the same clinical data. The two phenotype components of the mixture were identical to the linear model fit in the previous section. The results from both components are given in Table 2.6. The BIC for the mixture model was 270.03.

	Estimate	SD	p	Estimate	SD	p
Intercept	-0.1134	0.1705	0.5063	-0.2926	0.6634	0.6594
BMI	0.0033	0.0016	0.0353	0.0263	0.0079	0.0009
Urine Output	-0.0000	0.0000	0.1374	-0.0003	0.0001	0.0083
Fluid Given	-0.0000	0.0000	0.6474	0.0000	0.0000	0.3639
Baseline SCr	-0.0876	0.1382	0.5268	0.6862	0.5583	0.2196
Age	0.0021	0.0021	0.3187	0.0060	0.0085	0.4800
Length of Surgery	0.0004	0.0001	0.0009	0.0007	0.0005	0.1793
Maximum Lactate Level	0.0029	0.0057	0.6164	-0.0377	0.0283	0.1838
Cross Clamp Time	-0.0002	0.0002	0.3223	0.0022	0.0008	0.0045
Baseline - Mean Diastolic	-0.0008	0.0006	0.2411	-0.0029	0.0027	0.2847
Pulse Pressure	0.0005	0.0003	0.1105	0.0016	0.0014	0.2374
Bypass Time	0.0000	0.0001	0.8849	0.0010	0.0006	0.0857
Hespan	-0.0000	0.0000	0.9372	0.0004	0.0002	0.1138
Hypertension	-0.0093	0.0156	0.5530	0.1298	0.0655	0.0481
Diabetes	-0.0164	0.0180	0.3629	-0.1488	0.0973	0.1269
CHF	-0.0015	0.0180	0.9316	0.1132	0.0875	0.1961
COPD	-0.0335	0.0304	0.2706	-0.1496	0.1367	0.2744
HB	-0.0073	0.0046	0.1149	-0.0835	0.0237	0.0005
Baseline SCr*Age	0.0003	0.0021	0.8692	-0.0066	0.0077	0.3915

Table 2.6: Results from the mixture model applied to clinical data. The estimates on the left represent the null-subpopulation and the right represent the valid surrogate subpopulation.

2.6.3 Comparing the Linear vs. Mixture Model Approach

The mixture approach was heavily favored by BIC with a difference of 255.35 between the two models. This is evidence that the phenotypes were sufficiently different to allow the EM algorithm to seperate them reliably. This is further evidenced by examining the distribution of posterior probabilities produced by the EM algorithm, see Figure 2.5.

The difference in fit between the linear and mixture model approaches is also evident when examining the calibration plots that compare the linear model fit to that of the mixture model with posterior probability assignments, see Figure 2.6.

It is clear that the mixture model provides superior fit, and the hallmarks of having successfully identified the phenotypes defined by the partial surrogate are present. Thus, we expect that the component of the mixture representing the valid surrogate subpopulation contains legitimate hypothesis tests of association between the proposed risk factors and

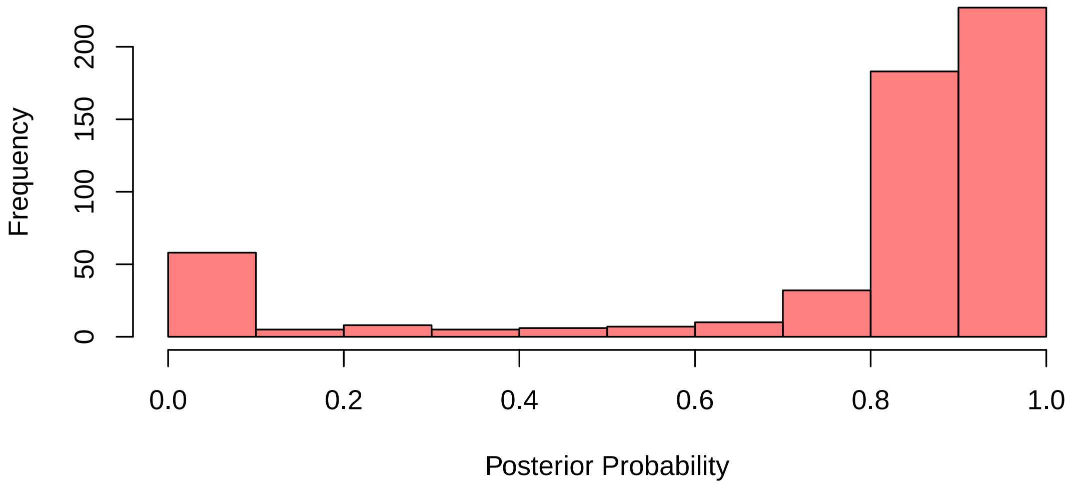


Figure 2.5: Distribution of posterior probabilities of being assigned to the null-subpopulation resulting from the mixture model applied to the clinical dataset.

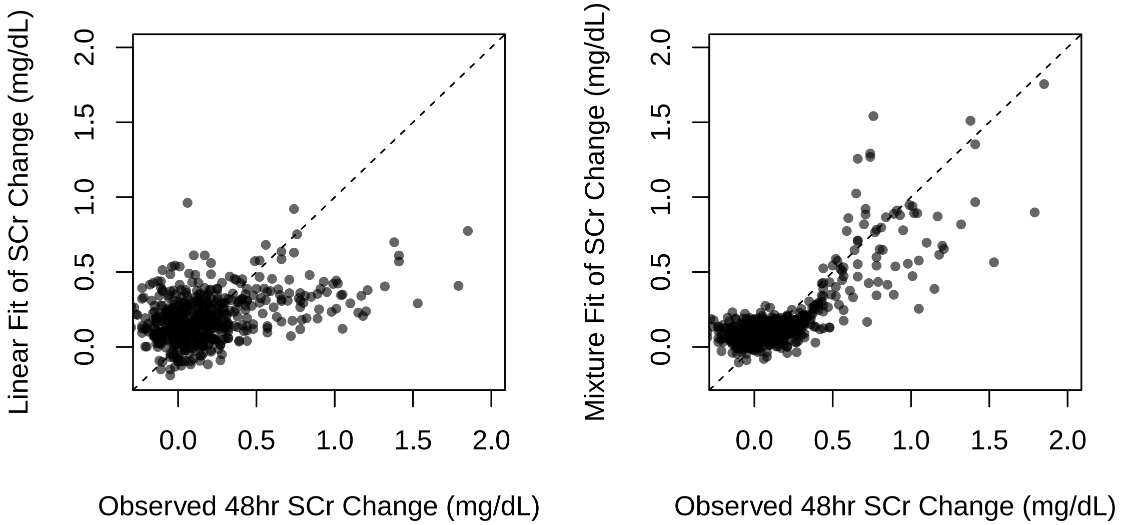


Figure 2.6: Calibration plots for the linear (left) and mixture (right) models applied to the clinical dataset.

nephron damage because we know suppose the Prentice criteria are satisfied in that subpopulation. However, careful consideration must be taken to determine to what surgical population these results are generalizable. In the partial surrogate's phenotypes are defined by a threshold effect, i.e. if the amount of nephron damage you receive is sufficient to exhaust your functional reserve the surrogate measure will be valid. As a consequence of this thresholding patients demonstrating the non-null phenotype are sampled from those who receive substantive damage and it is not clear that the identified associations will apply to more mild incidents of damage. These hypothesis tests are therefore legitimate associations amongst those who develop AKI at the least. Further work is needed to show that these associations can be valuable for predicting sub-clinical AKI, i.e. those sustaining nephron damage that are part of the null-phenotype. We will revisit this idea later in the chapter.

2.6.4 Prediction of Short Term Outcomes

Despite having set out to acheive legitimate hypothesis testing with respect to nephron damage precisely because of serum creatinine's deficiency as a biomarker, there is no other standard of any kind to benchmark the performance of the mixture model approach in the acute perioperative period. Accordingly, if the mixture model were capable of predicting which patients were likely to develop AKI in the first 48 hours directly after surgery this would still represent a clinical utility of the model beyond its ability to provide appropriate hypothesis tests.

Obtaining predictions from a latent variable model does not always yield satisfying results even when the model fits well. The mixture gives two predictions for each person, one for each phenotype. Therefore we must predict to which phenotype a patient likely belongs, or at least their probability of being of a particular phenotype. Having predicted phenotype probabilities in hand, we can implement an empirical Bayes approach to prediction, i.e. averaging the predictions according to their predicted probability of representing the correct phenotype.

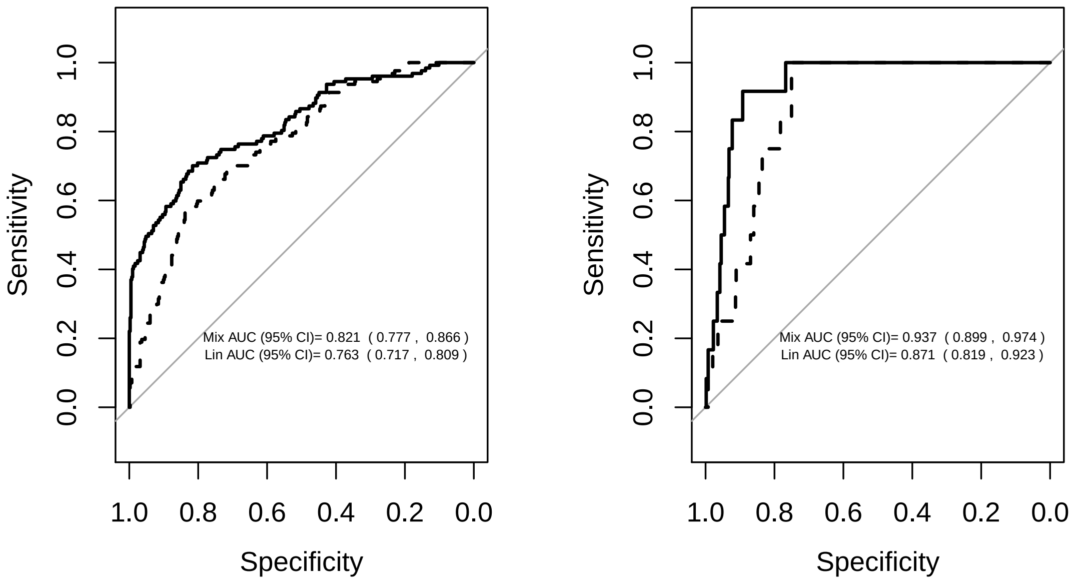


Figure 2.7: ROC curves for KDIGO1 (left) and KDIGO2 (right).

A regression-type support vector machine with a radial kernel was fit to the posterior probabilities resulting from the EM algorithm containing the same covariates used to fit the mixture model. This model was used to provide predicted probabilities and mixture predictions were calculated as discussed above. The predictive ability of this approach was then assessed via mean square error (MSE) and area under the receiver operating characteric curve for various cutoffs determined to be clinically meaningful.

Of the most widely used clinical definitions of perioperative AKI the most current is the KDIGO criteria (14). This criteria is staged but only the first two stages are represented in this dataset. The ROC curves for these two clinical criteria are represented in Figure 2.7. The p-values of DeLong's test for the difference in AUC were 5×10^{-4} and 0.0013 for KDIGO I and KDIGO II respectively.

Perhaps the most common measure used to assess prediction accuracy is mean square prediction error. The apparent mean square prediction error for the two models was 0.123 and 0.112 for the linear and mixture models respectively. This is a relative reduction in MSE of 8.9% by using the mixture approach. The MSE difference was validated in 2000 bootstrap replications whose results are presented in Table 2.7.

Mixture	Linear	p-value
0.115(0.099, 0.13)	0.129(0.121, 0.133)	0.034

Table 2.7: MSE results from 2000 bootstrap replications to validate the model comparison.

By utilizing the support vector machine model to get estimates of the probability of the latent variable, it is possible to use the mixture model approach to achieve more accurate predictions of the maximum serum creatinine change within 48 hours than the linear model is capable of producing. This benefit is due to the support vector machine's ability to model abstract relationships to get accurate predictions of the fitted responsibilities from the EM algorithm. These predictions allow the calculation of an estimated empirical Bayes posterior probability and corresponding estimate of serum creatinine change.

2.6.5 Predicting Long-Term Outcomes

In a previous section it was suggested that further study would be necessary to understand the population to which associations estimated by the mixture model approach would be generalizable. In particular, a different approach would be required to demonstrate that the mixture approach was valid among patients with a more minor degree of nephron damage. As discussed in the previous section, there is no gold standard in the perioperative period to which the mixture predictions could be compared to support this hypothesis. Minor injuries could however lead to more long term renal adverse events after the patient has been discharged from the hospital. The general hypothesis would be that the portion of the mixture model that represents the valid surrogate phenotype's predictions could be extrapolated to lesser renal injuries. Although it is unlikely that this model's predictions are well calibrated, it is possible that these predictions maintain a degree of discrimination that can be utilized for detecting subclinical injury.

Unfortunately among the 541 clinical trial patients in the data used in the last section only 30 of them had any longer term followup data that could be utilized for a longer term

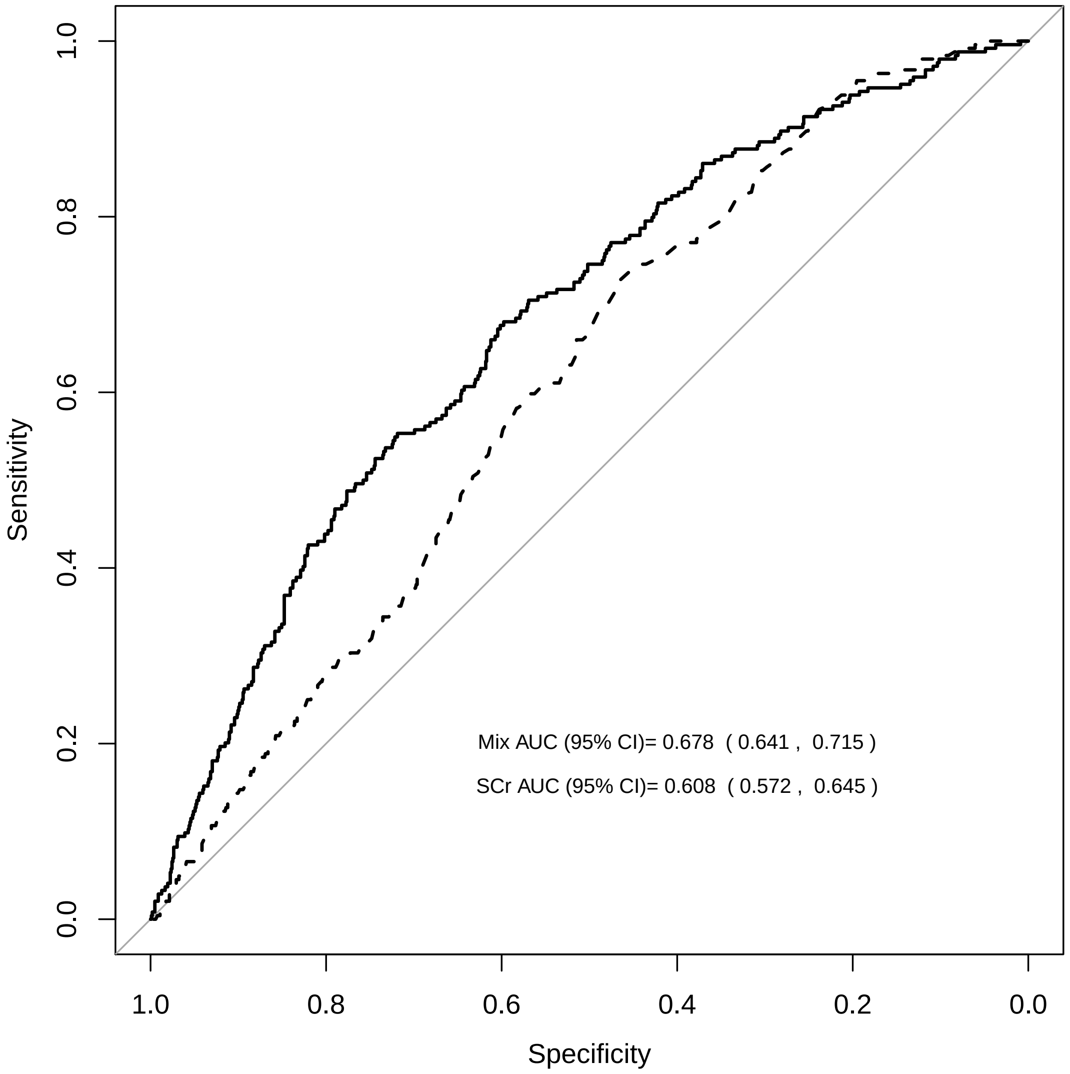


Figure 2.8: ROC curve for predicting 90-day eGFR decrease exceeding 20 for the mixture model (solid) versus perioperative serum creatinine change (dashed).

analysis. In an attempt to gain more precision in estimating the performance of the mixture model, a larger data set of cardiac surgery patients is necessary. This data set is observational in nature and as a consequence has major issues with missing data, particularly lab values and intraoperative records. These issues will require a modification of the included variables to exclude variables that are extensively missing. In addition, the patients there is no compelling reason to believe the patients that have longer-term data available (n = 1268) are a random sample, so there is the potential that the outcome variable is not missing at random. With all these limitations in mind, a comparison was made as to how well the mixture model predicts 90-day decreased renal function as compared to perioperative serum creatinine change.

The most widely utilized measure of chronic kidney function is called estimated glomerular filtration rate (eGFR). Although also a serum creatinine based measure, because it is calculated at steady-state conditions it does not suffer from the problems induced on serum creatinine by the activation of renal functional reserve in cases of acute injury. Decreased eGFR is indicative of progression of chronic renal disease.

The discrimination of the mixture model vs perioperative serum creatinine change was first assessed via AUC. A ROC curve was constructed for the prediction of 90-day eGFR decrease greater than or equal to 20 ml/min/m^2, see Figure 2.8. Delong's test for the difference in AUCs resulted in a p-value of 0.002. In order to demonstrate that this measure is not cutoff specific, the two candidate measures were also compared using spearman's rank correlation. The correlations with 90-day eGFR change were 0.305 and 0.231 for the mixture approach and serum creatinine change respectively. A permutation test was done with 5000 replications to compare these correlations resulting in a p-value of 0.033.

These results suggest that the mixture model approach was able to predict injuries better than perioperative serum creatinine change. This is evidence that the mixture approach is able to discern biological changes that serum creatinine cannot.

2.7 Manuscript 1: Clinical Presentation of AKI results

2.7.1 Abstract

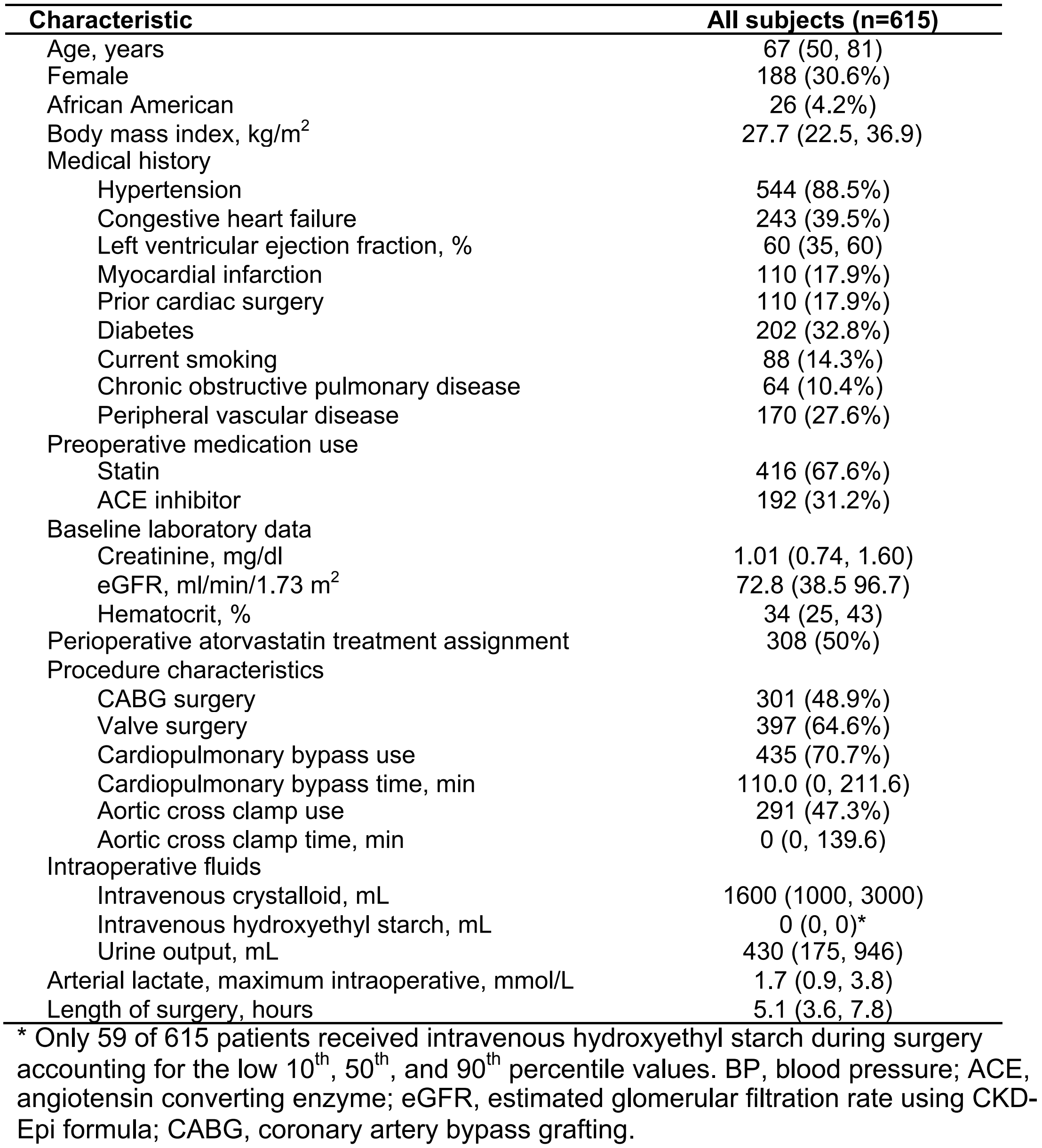
Acute kidney injury (AKI), a serious adverse event following cardiac surgery, is diagnosed based on postoperative serum creatinine change. Models for predicting the risk for AKI have not consistently performed well, likely due to the omission of clinically important, but practically unmeasurable, variables. We hypothesized that a latent variable mixture model of postoperative serum creatinine change following cardiac surgery would partially account for these unmeasured factors and therefore increase power to identify risk factors of AKI and improve predictive accuracy compared to a traditional linear model. We constructed a two-component latent variable mixture model and a linear model using data from a prospective, 653-subject randomized clinical trial of AKI following cardiac surgery and included established AKI risk factors and covariates known to affect serum creatinine values. The latent variable mixture model demonstrated superior fit (likelihood ratio of 6.68x1071) and enhanced discrimination (permutation test of Spearmans correlation coefficients, p <0.001) compared to the linear model. The latent variable mixture model was 94% (-13% to 1132%) more powerful (median [range]) at identifying risk factors as the linear model, and demonstrated increased ability to predict postoperative change in serum creatinine (relative mean square error reduction of 6.8%). Incorporation of latent variable mixture modeling into AKI research will allow clinicians and investigators to account for clinically meaningful patient heterogeneity resulting from unmeasured variables, and therefore provide improved ability to examine risk factors, measure mechanisms and mediators of kidney injury, and more accurately predict AKI.

2.7.2 Introduction

The diagnosis of acute kidney injury (AKI) relies primarily on changes in serum creatinine concentrations (SCr).[13] Changes in serum creatinine, however, can be insensitive and nonspecific for renal injury [4,5] due to unmeasured confounders such as changes in creatinine production, myocyte injury, intravenous fluid administration, and renal functional reserve.[6,7] Many of these unmeasured confounders, also known as latent variables,[8] represent an important source of patient heterogeneity with respect to how and if AKI manifests, but are clinically impractical or impossible to measure. Failure to account for factors like these decreases power to identify risk factors for AKI and hinders accurate prediction of postoperative AKI. Latent variable mixture modeling improves the ability to assess the associations between independent variables and an outcome by accounting for the effect of a latent variable. The model uses measured covariates to empirically stratify a cohort into subpopulations of patients, represented by a latent variable, which are more homogenous than the total cohort. These subpopulations are represented by component models which can be combined to form a comprehensive model to represent the entire cohort[9]. We hypothesized that a latent variable mixture model would increase power to identify significant risk factors for AKI and improve accuracy in predicting a patients postoperative SCr compared to a traditional linear model. To test this hypothesis, we built a traditional linear model and a two-component latent variable mixture model to predict SCr in a well-phenotyped clinical trial of AKI following cardiac surgery and compared the models goodness-of-fit, power to identify established AKI risk factors, discrimination, and prediction of 48-hour postoperative SCr.

2.7.3 Results

Subject Characteristics and AKI Six hundred fourteen patients comprised the study cohort. The cohort was primarily Caucasian, and one third of patients were female (Table 1). Half of the patients received coronary artery bypass surgery, two-thirds valve replacement or repair, and three quarters of surgeries were performed with the use of cardiopulmonary bypass. One hundred thirty five patients (22.1%) developed KDIGO AKI. One hundred and nineteen of these patients met the 0.3 mg/dL increase within 48-hour criterion, 72 the



Characteristic	All subjects (n=615)
Age, years	67 (50, 81)
Female	188 (30.6%)
African American	26 (4.2%)
Body mass index, kg/m^2	27.7 (22.5, 36.9)
Medical history	
Hypertension	544 (88.5%)
Congestive heart failure	243 (39.5%)
Left ventricular ejection fraction, %	60 (35, 60)
Myocardial infarction	110 (17.9%)
Prior cardiac surgery	110 (17.9%)
Diabetes	202 (32.8%)
Current smoking	88 (14.3%)
Chronic obstructive pulmonary disease	64 (10.4%)
Peripheral vascular disease	170 (27.6%)
Preoperative medication use	
Statin	416 (67.6%)
ACE inhibitor	192 (31.2%)
Baseline laboratory data	
Creatinine, mg/dl	1.01 (0.74, 1.60)
eGFR, ml/min/1.73 m^2	72.8 (38.5 96.7)
Hematocrit, %	34 (25, 43)
Perioperative atorvastatin treatment assignment	308 (50%)
Procedure characteristics	
CABG surgery	301 (48.9%)
Valve surgery	397 (64.6%)
Cardiopulmonary bypass use	435 (70.7%)
Cardiopulmonary bypass time, min	110.0 (0, 211.6)
Aortic cross clamp use	291 (47.3%)
Aortic cross clamp time, min	0 (0, 139.6)
Intraoperative fluids	
Intravenous crystalloid, mL	1600 (1000, 3000)
Intravenous hydroxyethyl starch, mL	0 (0, 0)*
Urine output, mL	430 (175, 946)
Arterial lactate, maximum intraoperative, mmol/L	1.7 (0.9, 3.8)
Length of surgery, hours	5.1 (3.6, 7.8)

* Only 59 of 615 patients received intravenous hydroxyethyl starch during surgery accounting for the low 10th, 50th, and 90th percentile values. BP, blood pressure; ACE, angiotensin converting enzyme; eGFR, estimated glomerular filtration rate using CKD-Epi formula; CABG, coronary artery bypass grafting.

Table 2.8: Cohort characteristics. Binary characteristics are reported as n (%) and continuous characteristics as median (10th percentile, 90th percentile).

50% increase within 7-days criterion, and 60 both. Twenty-six patients (4.2% of the total cohort) developed KDIGO stage II or III AKI, 5 of whom required postoperative renal replacement therapy.

The two-component latent variable mixture model identified two distinct subpopulations of patients indicating the existence of a latent variable. At the completion of model fitting, 13% of patients had >50% probability of being in subpopulation 1, and 87% of patients had <50% probability of being in subpopulation 1 (i.e., 87% of patients had >50% probability of being in subpopulation 2 (Figure 1)). If patients with a >50% probability of being in subpopulation 1 are assigned to subpopulation 1 and patients with >50% probability of being in subpopulation 2 are assigned to subpopulation 2, then in general subpopulation 1 tended to be older, with a greater prevalence of hypertension, diabetes, and congestive heart failure, and a lower baseline eGFR (Supplemental Table).

Latent Variable Mixture Model Subpopulation Assignments The two-component latent variable mixture model created two theoretical subpopulations based on the hypothesized existence of a latent variable. At the completion of model fitting, 13% of patients had >50% probability of being in theoretical subpopulation 1, and 87% of patients had <50% probability of being in subpopulation 1 (i.e., 87% of patients had >50% probability of being in subpopulation 2). Figure 1 shows the two distinct subpopulations of patients identified by the mixture model.

Model fit The latent variable mixture model demonstrated superior goodness-of-fit throughout the range of predicted SCr (Figure 2), resulting in a Bayesian Information Criteria (BIC) value of 140 for the latent variable mixture model compared to 349 for the linear model. These BIC values represent a $6.66x10^{7}1$ times increased likelihood of the latent variable mixture model providing superior fit compared to the linear model.

AKI risk factor identification and estimation accuracy The latent variable mixture model identified a significant association between 14 of the 16 established AKI risk factors included as covariates and maximum 48-hour SCr, while the linear model demonstrated a

Risk Factor	Linear Model	Latent Variable Mixture Model	
		Subpopulation 1	Subpopulation 2
Age (per 10 years)	0.037 (-0.044, 0.118)	0.040 (-0.021, 0.101)	0.042 (0.016, 0.068) †
BMI (per 5 kg/m^2)	0.040 (0.013, 0.068)†	0.107 (0.081, 0.132)‡	0.012 (0.005, 0.019)‡
History of hypertension	0.005 (-0.046, 0.056)	0.080 (0.002, 0.158)*	-0.017 (-0.031, -0.003)*
History of diabetes	-0.024 (-0.082, 0.035)	-0.123 (-0.177, -0.068)‡	-0.007 (-0.021, 0.008)
Baseline pulse pressure (per 10 mmHg)	0.003 (-0.009, 0.015)	-0.019 (-0.035, -0.003)*	0.004 (7.2e-5, 0.008)*
Baseline SCr (per mg/dL)	0.203 (-0.309, 0.715)	0.054 (-0.217, 0.326)	0.158 (-0.023, 0.339)
Baseline SCr:age interaction	-0.001 (-0.008, 0.007)	0.002 (-0.003, 0.007)	-0.001 (-0.003, 0.002)
Baseline eGFR (per 30 mL/min/1.73 m^2)	0.081 (-0.012, 0.174)	0.045 (-0.066, 0.156)	0.099 (0.051, 0.147)‡
Baseline hematocrit (per %)	-0.010 (-0.016, -0.005)‡	-0.034 (-0.040, -0.028)‡	-0.003 (-0.005, -0.001)‡
Cardiopulmonary bypass time (per hour)	0.006 (-0.018, 0.030)	-0.072 (-0.108, -0.036)‡	0.012 (2.0e-4, 0.024)†
Aortic cross clamp time (per hour)	0.036 (0.001, 0.072)*	0.156 (0.120, 0.192)‡	-0.006 (-0.018, 0.006)
Intra-operative hydroxyethyl starch volume (per L)	0.200 (0.000, 0.400)	0.300 (0.100, 0.500)†	0.000 (-0.056, 0.094)
Intraoperative urine output (per L)	-0.100 (-0.200, -0.048)†	-0.300 (-0.400, -0.200)‡	-0.100 (-0.094, -0.016)‡
Mean intra-operative MAP adjusted for baseline MAP (per 10 mmHg)	0.023 (0.003, 0.042)*	0.064 (0.042, 0.086)‡	0.005 (0.001, 0.009)*
Maximum intra-operative lactate (per mmol/L)	0.004 (-0.021, 0.028)	-0.009 (-0.033, 0.016)	0.013 (0.007, 0.019)‡
Length of surgery (per hour)	0.034 (0.009, 0.059)†	0.113 (0.086, 0.140)‡	0.027 (0.021, 0.034)‡

*p<0.05, †p<0.01, ‡p<0.001; BMI, body mass index; eGFR, estimated glomerular filtration rate

using CKD-Epi formula, SCr, serum creatinine concentration; MAP, mean arterial blood pressure

Table 2.9: Associations between established AKI risk factor covariates and maximum 48-hour serum creatinine change from baseline using a linear model and each subpopulation of a two-component latent variable mixture model. For example, an increase of ten years in age is associated with a 0.037 increase in 48-hour postoperative change in serum creatinine concentration (SCr) in the linear model, and a past medical history of hypertension was associated with a 0.080 increased in 48-hour SCr in the subpopulation 1 component model. Ninety-five percent confidence intervals are listed after each covariate coefficient estimate.

significant association between 6 of the 16 established AKI risk factors and maximum 48-hour SCr (Table 2). Post hoc relative power calculations showed that the latent variable mixture model had greater power to identify established risk factors as significant for 15 of the sixteen covariates considered compared to the linear model (sign test, p <0.001). The latent variable mixture model exhibited 94% (-13% to 1132%) more power (median [range]) to identify established risk factors as having a statistically significant association with 48-hour SCr as the linear model.

A quantile-quantile (Q-Q) plot revealed that the latent variable mixture model deviated less from the line of best fit than the linear model (Figure 3), demonstrating that the latent variable mixture model better fulfilled the linear regression requirement of normally distributed errors. This signifies that the latent variable mixture model has an improved ability to accurately assess associations between patient characteristics and postoperative SCr compared to the linear model.

Model Discrimination and Prediction of 48-hour postoperative SCr The latent variable mixture model demonstrated superior discrimination for predicted SCr compared to the linear model (Permutation test of Spearmans correlation coefficients, p <0.001). The relative mean squared error reduction for the latent variable mixture model comparative to the linear model was 6.8%, meaning that the latent variable mixture model predicted 48-hour postoperative SCr 6.8% more accurately.

2.7.4 Discussion

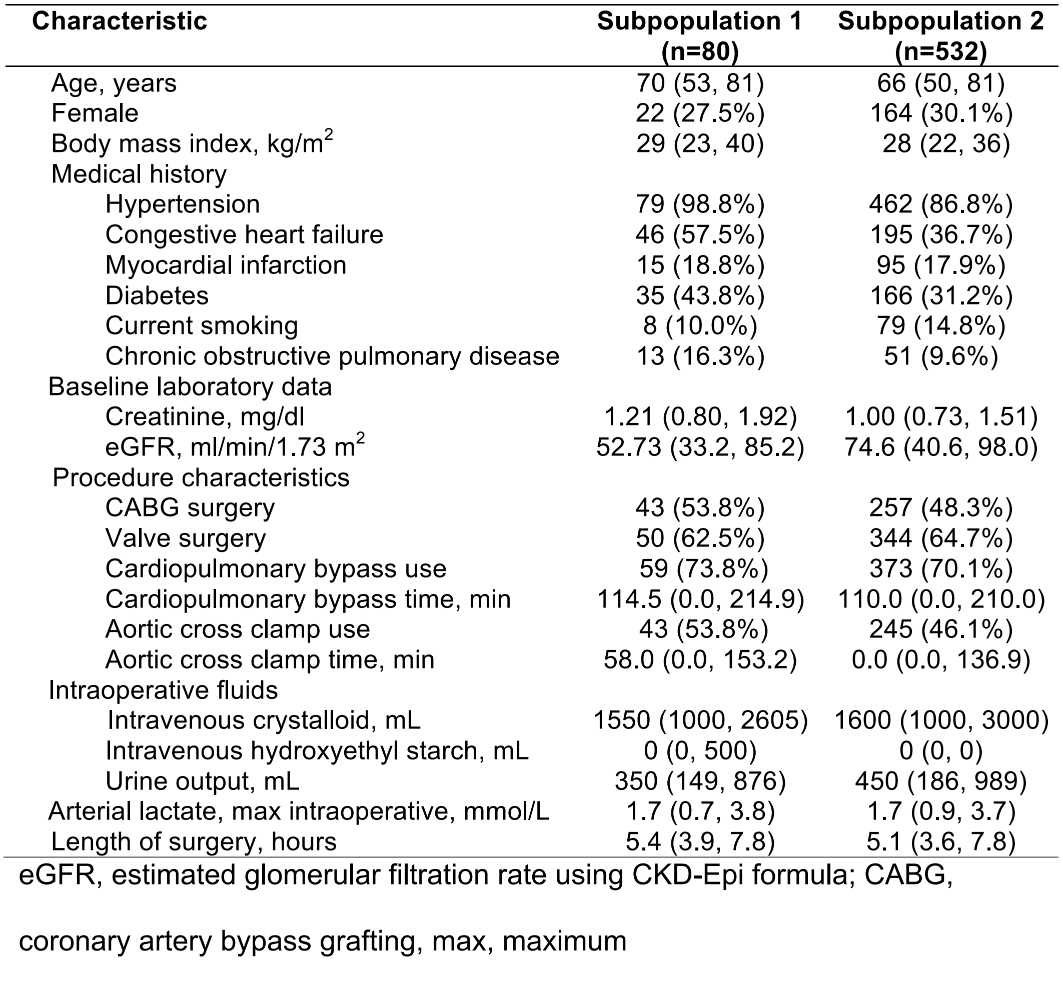
In this study of perioperative AKI, a latent variable mixture model had markedly more power to identify established risk factors for AKI and improved ability to predict a patients postoperative SCr than a traditional linear model. These benefits were likely due to superior goodness-of-fit, improved accuracy of covariate coefficient estimation, and enhanced discrimination of predicted postoperative SCr. Latent variable mixture modeling may offer substantial benefits to the study of AKI, and future studies that seek to isolate risk factors for AKI, measure mechanisms of AKI, test therapies for AKI, or seek to predict AKI in clinical cohorts should consider using this methodology. The improvement of AKI modeling with the latent variable mixture modeling technique indicates that substantial heterogeneity exists within the perioperative AKI population that is not accounted for by observed covariates, and that reliance on traditional linear modeling techniques which inherently assume observed covariates are the only relevant covariates obscures this heterogeneity. This unaccounted for patient heterogeneity within the AKI population may explain why numerous AKI prevention and intervention trials have failed to demonstrate efficacy despite promising preclinical trials. While new to studies of AKI, latent variable mixture modeling is an established statistical methodology to account for patient heterogeneity in other clinical domains. It has long been used in psychology and genetics research,[10,11] and more recently in oncology. For example, the use of latent variable mixture modeling to model small cell lung cancer growth dynamics from serum biomarker data has improved the prediction of treatment outcomes and decreased reliance on sequential imaging. [12] In acute lung injury, a latent variable mixture modeling technique recently identified patient phenotypes associated with differential treatment effects of high versus low positive end expiratory pressure where traditional modeling had failed[13]. Identification of latent variable subpopulations in patients at risk for AKI may also lead to the identification of subpopulation-specific treatment benefits, enhanced risk stratification, and improved prediction of long-term outcomes. In the current study, the latent variable mixture model displayed greater power to identify established risk factors for AKI. This improvement results in increased power to identify and characterize novel candidate risk factors, including baseline characteristics, intraoperative exposures, perioperative biomarkers, and patient management techniques that could be modified to reduce AKI. A latent variable mixture modeling assessment of candidate factors will increase discernment of their effects on AKI and benefit the search for other non-latent, modifiable AKI risk factors, particularly in modestly sized patient cohorts where power may be low. Development of the latent variable mixture model does not itself identify the latent variable or binomial pattern of variables, but can suggest potential candidates including renal functional reserve, genetic polymorphisms, clusters of disease exposure, fluid management strategies, or surgical treatments. For example, renal functional reserve is a potentially source of heterogeneity in susceptibility that leads to variation in the manifestation of AKI across patients. [7,1419] In our study, older age and higher comorbidity burden (e.g. diabetes, hypertension) is potentially consistent with a population with less renal reserve compared to subpopulation 2 in whom traditional risk modeling performed less well [7,20,21]. In the former, a potential lack of renal reserve might explain the larger model coefficients associated with established AKI risk factors such as history of hypertension and diabetes, BMI, baseline hematocrit, aortic cross clamp duration, and length of surgery. In contrast, the potential presence of renal functional reserve might contribute to smaller, and frequently statistically insignificant, model coefficients for established AKI risk factors in subpopulation 2. The latter subpopulation might represent patients in whom sensitive AKI biomarkers may better predict the potential long-term impact of AKI than currently emphasized risk factors. Irrespective of the identity of the latent variable, our results indicate that latent variable mixture modeling can identify subpopulations of patients that may be used to enrich outcomes in clinical trials, target monitoring and interventions, and shed novel insight into the pathophysiology of AKI. Strengths of this study include the use of high-quality unbiased data collected as part of a prospective clinical trial with little to no missing data. We also retained serum creatinine as a continuous variable to enhance AKI discrimination and prediction [22,23]. At the same time we acknowledge potential limitations. We did not evaluate latent variable mixture models with more than two subpopulations or perform latent class analysis to empirically determine the number of subpopulations to model. Given the goal of comparing latent variable mixture modeling to traditional linear modeling techniques, we selected the simplest latent variable mixture model for this initial assessment. We observed dramatic results, but increased latent variable flexibility could further improve AKI modeling. A second limitation was the small number of patients that developed moderate or severe AKI (100% or 200% SCr KDIGO stage II/III), which limited our power to compare latent variable mixture modeling to linear modeling techniques with high precision in patients with moderate or severe AKI. A majority of patients that develop postoperative AKI, however, develop mild AKI, and this outcome remains associated with major short and long-term morbidity [2426]. In conclusion, a latent variable mixture model increased power to identify established AKI risk factors, more accurately ranked the severity of patients 48-hour SCr, and more accurately predicted 48-hour postoperative SCr compared to a linear model. Latent variable mixture modeling may improve clinicians ability to identify novel risk factors and advance the understanding of AKI pathophysiology. Employment of this technique could also advance pre-operative AKI risk stratification and provide opportunities to further phenotype and target higher risk patient subpopulations with specific monitoring, preventative strategies, and treatments. Latent variable mixture modeling may provide a powerful technique to advance the study of AKI.

2.7.5 Methods

Patient Sample After Institutional Review Board approval, we collected data from a 653-subject prospective clinical trial of perioperative statin use to prevent AKI following cardiac surgery conducted at a large academic medical center from 2009-2014. The study was conducted according to the Declaration of Helsinki. Patients were eligible to participate in the trial if they were scheduled for elective coronary artery bypass grafting, valve surgery, or ascending aortic surgery requiring thoracotomy or sternotomy. Patients receiving preoperative renal replacement therapy, with liver dysfunction, acute coronary syndrome, pregnancy, current CYP3A4 inhibitor use, and a history of kidney transplant or statin intolerance were ineligible to participate. Six hundred fifty-three patients provided written informed consent. Thirty-eight patients were excluded for failing inclusion criteria or withdrew for personal reasons prior to study initiation, and one patient that completed the study received hemodialysis on postoperative day one and was excluded from 48-hour SCr model development since this patients SCr no longer reflected renal injury or function. Thus 614 patients were included. No significant association between perioperative statin use and postoperative AKI was demonstrated in the clinical trial [27]. Modeling AKI We chose maximum SCr from baseline to postoperative day 2 to model AKI because serum creatinine is the most common and best characterized marker of renal injury, a 48-hour interval is consistent with current consensus guidelines for AKI diagnosis, and a continuous scale rather than a binomial threshold for AKI preserves the measurement of AKI severity and provides the best opportunity to ascertain differences between linear and latent variable mixture modeling techniques. Baseline serum creatinine concentration was defined as the most recent preoperative creatinine measurement and was measured in inpatients on the morning of surgery and within a week prior to surgery in outpatients. Postoperative serum creatinine concentrations were measured at 2:00 am daily throughout hospitalization. We selected model covariates a priori based on established predictors of post-cardiac surgery AKI and factors known to affect serum creatinine production or dilution [6,9,2831]. Including well-established risk factors for AKI facilitates comparison of each models ability to identify significant AKI risk factors for the prediction of SCr. Selected covariates were identical for both the linear model and the latent variable mixture model and included age, body mass index (BMI), baseline glomerular filtration rate estimated using the CKD-EPI formula (eGFR),[32] baseline serum creatinine, agebaseline serum creatinine interaction term, baseline hematocrit, presence of diabetes, presence of hypertension, duration of surgery, baseline pulse pressure, volume of hydroxyethyl starch administered during surgery, volume of urine output during surgery, duration of cardiopulmonary bypass, duration of aortic cross clamp, maximum intraoperative arterial lactate concentration, and average intraoperative mean blood pressure adjusted for baseline mean blood pressure. Dataset completion was excellent (100% of all serum creatinine data were complete; >99% of all covariate data were complete).

Model development A linear model and a two-component latent variable mixture model were each fit to the maximum SCr from baseline over the first 48 postoperative hours. The latent variable mixture model is composed of two traditional linear models, known as component models. Each component model represents a subpopulation of patients formed by the latent variable. During fitting, the mixture model agnostically identifies two distinctive subpopulations based on covariate patterns with respect to observed 48-hour postoperative SCr. Given that there is uncertainty regarding individual patient subpopulation membership (i.e., subpopulation membership is determined by each patients unknown latent variable status, 0 or 1), a probability of being in each subpopulation is initially randomly assigned to each patient and then refined during the iterative model fitting process until convergence criteria are met. Therefore at the conclusion of model fitting, a patient whose covariate pattern is very consistent with subpopulation 1, for example, may be assigned a 90% probability of subpopulation 1 membership and a 10% probability of subpopulation 2 membership. In this way, each patients data may contribute to both component models, improving overall model fit. Each component model represents a data-identified patient subpopulation. If two distinct subpopulations are not identified during the fitting process, the first component model would become identical to the traditional linear model and the coefficients for all the covariates of the second component model would be assigned a value of zero. After completion of model fitting, we developed a support vector machine algorithm to predict patient subpopulation allocation probabilities based on covariate patterns but independent of observed SCr. This enables prediction of SCr using the latent variable mixture model and allows us to compare SCr prediction between latent variable mixture and linear models.

Statistical analyses Patient characteristics were summarized with the 50th (10th, 90th) percentiles for continuous variables and percentages for categorical variables. To evaluate the latent variable mixture model relative to the linear model, we compared model: 1) goodness-of-fit, 2) average power to identify established risk factors for AKI, 3) discrimination (ability to rank subjects in order of predicted SCr), and 4) accuracy to predict maximum 48-hour postoperative SCr. Goodness-of-fit was assessed with calibration plots and R^2 calculations of predicted SCr versus observed SCr for the latent variable mixture and linear models. To account for the increased flexibility of the latent variable mixture model with respect to differential model fit, BIC values were calculated for each model and compared using a relative likelihood calculation. The ability to identify AKI risk factors was assessed using a post hoc calculation of each models power to identify established risk factors as significant. For this calculation, 5000 new datasets were generated from our original dataset using standard parametric bootstrapping techniques, and both the latent variable mixture model and the linear model were refit in each new dataset. This produced a set of new risk factor coefficients and associated p-values for each model. Using these sets of new model coefficients, individual risk factor identification power comparisons between the two models were performed, taking our original fitted model coefficients as the power calculations alternative hypotheses. A sign test was used to determine the significance of the power comparison between the two models. Additionally, Q-Q plots were used to assess the normalcy of each models errors in order to compare each models covariate coefficient accuracy. Model discrimination was evaluated with a permutation test of each models Spearmans correlation coefficients between predicted and observed SCr. To evaluate prediction of maximum 48-hour postoperative SCr, we compared the average of the square of the difference between the predicted and true SCr (i.e., mean squared error relative difference [(predicted SCr true SCr)2]) [33]. Models were bootstrapped with 200 replicates to assess for over-fitting and provide internal validation. Statistical analyses were performed in R (version 3.2.0, R Foundation, http://www.r-project.org) and included pROC and flexmix packages.



Characteristic	Subpopulation 1 (n=80)	Subpopulation 2 (n=532)
Age, years	70 (53, 81)	66 (50, 81)
Female	22 (27.5%)	164 (30.1%)
Body mass index, kg/m^2	29 (23, 40)	28 (22, 36)
Medical history		
Hypertension	79 (98.8%)	462 (86.8%)
Congestive heart failure	46 (57.5%)	195 (36.7%)
Myocardial infarction	15 (18.8%)	95 (17.9%)
Diabetes	35 (43.8%)	166 (31.2%)
Current smoking	8 (10.0%)	79 (14.8%)
Chronic obstructive pulmonary disease	13 (16.3%)	51 (9.6%)
Baseline laboratory data		
Creatinine, mg/dl	1.21 (0.80, 1.92)	1.00 (0.73, 1.51)
eGFR, ml/min/1.73 m^2	52.73 (33.2, 85.2)	74.6 (40.6, 98.0)
Procedure characteristics		
CABG surgery	43 (53.8%)	257 (48.3%)
Valve surgery	50 (62.5%)	344 (64.7%)
Cardiopulmonary bypass use	59 (73.8%)	373 (70.1%)
Cardiopulmonary bypass time, min	114.5 (0.0, 214.9)	110.0 (0.0, 210.0)
Aortic cross clamp use	43 (53.8%)	245 (46.1%)
Aortic cross clamp time, min	58.0 (0.0, 153.2)	0.0 (0.0, 136.9)
Intraoperative fluids		
Intravenous crystalloid, mL	1550 (1000, 2605)	1600 (1000, 3000)
Intravenous hydroxyethyl starch, mL	0 (0, 500)	0 (0, 0)
Urine output, mL	350 (149, 876)	450 (186, 989)
Arterial lactate, max intraoperative, mmol/L	1.7 (0.7, 3.8)	1.7 (0.9, 3.7)
Length of surgery, hours	5.4 (3.9, 7.8)	5.1 (3.6, 7.8)

eGFR, estimated glomerular filtration rate using CKD-Epi formula; CABG, coronary artery bypass grafting, max, maximum

Table 2.10: Latent variable mixture model subpopulation characteristics. Binary characteristics are reported as n (%) and continuous characteristics as median (10th percentile, 90th percentile).

References

1. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P: Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. Crit Care 8: R204212, 2004

2. KDIGO AKI guideline work group. KDIGO Clinical Practice Guidelines for Acute Kidney Injury. Kidney Int Suppl 2: 1141, 2012

3. Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, Levin A, Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care 11: R31, 2007

4. Waikar SS, Betensky RA, Emerson SC, Bonventre JV: Imperfect gold standards for kidney injury biomarker evaluation. J Am Soc Nephrol 23: 1321, 2012

5. Waikar SS, Betensky RA, Bonventre JV: Creatinine as the gold standard for kidney injury biomarker studies? Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc 24: 32633265, 2009

6. Bellomo R, Kellum JA, Ronco C: Defining acute renal failure: physiological principles. Intensive Care Med 30: 3337, 2004

7. Sharma A, Mucino MJ, Ronco C: Renal functional reserve and renal recovery after acute kidney injury. Nephron Clin Pract 127: 94100, 2014

8. Schmiege SJ, Meek P, Bryan AD, Petersen H: Latent variable mixture modeling: a flexible statistical approach for identifying and classifying heterogeneity. Nurs Res 61: 204212, 2012

9. Berg KS, Stenseth R, Wahba A, Pleym H, Videm V: How can we best predict acute kidney injury following cardiac surgery?: a prospective observational study. Eur J Anaesthesiol 30: 704712, 2013

10. Bentley MJ, Lin H, Fernandez TV, Lee M, Yrigollen CM, Pakstis AJ, Katsovich L, Olds DL, Grigorenko EL, Leckman JF: Gene variants associated with antisocial behaviour:

a latent variable approach. J Child Psychol Psychiatry 54: 10741085, 2013

11. Xu MK, Gaysina D, Barnett JH, Scoriels L, van de Lagemaat LN, Wong A, Richards M, Croudace TJ, Jones PB, LHA genetics group: Psychometric precision in phenotype definition is a useful step in molecular genetic investigation of psychiatric disorders. Transl Psychiatry 5: e593, 2015

12. Buil-Bruna N, Lpez-Picazo J-M, Moreno-Jimnez M, Martn-Algarra S, Ribba B, Trocniz IF: A population pharmacodynamic model for lactate dehydrogenase and neuron specific enolase to predict tumor progression in small cell lung cancer patients. AAPS J 16: 609619, 2014

13. Calfee CS, Delucchi K, Parsons PE, Thompson BT, Ware LB, Matthay MA, NHLBI ARDS Network: Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. Lancet Respir Med 2: 611620, 2014

14. Bosch JP: Renal reserve: a functional view of glomerular filtration rate. Semin Nephrol 15: 381385, 1995

15. DeSanto NG, Anastasio P, Coppola S, Barba G, Jadanza A, Capasso G: Age-related changes in renal reserve and renal tubular function in healthy humans. Child Nephrol Urol 11: 3340, 1991

16. Epstein M: Aging and the kidney. J Am Soc Nephrol 7: 11061122, 1996

17. Rodrguez-Iturbe B, Herrera J, Garca R: Response to acute protein load in kidney donors and in apparently normal postacute glomerulonephritis patients: evidence for glomerular hyperfiltration. Lancet 2: 461464, 1985

18. Thomas DM, Coles GA, Williams JD: What does the renal reserve mean? Kidney Int 45: 411416, 1994

19. Friedman EA, Woredekal Y: Diabetic Nephropathy. In: Current diagnosis and treatment Nephrology and Hypertension, 1st ed., edited by Nissenson AR, Berns JS, Lerma EV, USA, McGraw-Hill, 2009, pp 483491

20. Haase M, Devarajan P, Haase-Fielitz A, Bellomo R, Cruz DN, Wagener G, Krawczeski

CD, Koyner JL, Murray P, Zappitelli M, Goldstein SL, Makris K, Ronco C, Martensson J, Martling C, Venge P, Siew E, Ware LB, Ikizler TA, Mertens PR: The outcome of neutrophil gelatinase-associated lipocalin-positive subclinical acute kidney injury: a multicenter pooled analysis of prospective studies. J Am Coll Cardiol 57: 17521761, 2011

21. Ronco C, Kellum JA, Haase M: Subclinical AKI is still AKI. Crit Care 16: 313316, 2012

22. Dawson NV, Weiss R: Dichotomizing continuous variables in statistical analysis: a practice to avoid. Med Decis Mak 32: 225226, 2012

23. Fitzsimons GJ: Death to Dichotomizing: Figure 1. J Consum Res 35: 58, 2008

24. Elmistekawy E, McDonald B, Hudson C, Ruel M, Mesana T, Chan V, Boodhwani M: Clinical Impact of Mild Acute Kidney Injury After Cardiac Surgery. Ann Thorac Surg 98: 815822, 2014

25. Gameiro J, Neves JB, Rodrigues N, Bekerman C, Melo MJ, Pereira M, Teixeira C, Mendes I, Jorge S, Rosa R, Lopes JA: Acute kidney injury, long-term renal function and mortality in patients undergoing major abdominal surgery: a cohort analysis. Clin Kidney J 9: 192200, 2016

26. Lassnigg A: Minimal Changes of Serum Creatinine Predict Prognosis in Patients after Cardiothoracic Surgery: A Prospective Cohort Study. J Am Soc Nephrol 15: 15971605, 2004

27. Billings FT, Hendricks PA, Schildcrout JS, Shi Y, Petracek MR, Byrne JG, Brown NJ: High-Dose Perioperative Atorvastatin and Acute Kidney Injury Following Cardiac Surgery: A Randomized Clinical Trial. JAMA 315: 877888, 2016

28. Harel Z, Chan CT: Predicting and preventing acute kidney injury after cardiac surgery. Curr Opin Nephrol Hypertens 17: 624628, 2008

29. Huen SC, Parikh CR: Predicting acute kidney injury after cardiac surgery: a systematic review. Ann Thorac Surg 93: 337347, 2012

30. Parolari A, Pesce LL, Pacini D, Mazzanti V, Salis S, Sciacovelli C, Rossi F, Ala-

manni F, Monzino Research Group on Cardiac Surgery Outcomes: Risk factors for perioperative acute kidney injury after adult cardiac surgery: role of perioperative management. Ann Thorac Surg 93: 584591, 2012

31. Kim WH, Lee SM, Choi JW, Kim EH, Lee JH, Jung JW, Ahn JH, Sung KI, Kim CS, Cho HS: Simplified clinical risk score to predict acute kidney injury after aortic surgery. J Cardiothorac Vasc Anesth 27: 11581166, 2013

32. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, Kusek JW, Eggers P, Van Lente F, Greene T, Coresh J, Chronic Kidney Disease Epidemiology Collaboration: A new equation to estimate glomerular filtration rate. Ann Intern Med 150: 604612, 2009

33. James G, Witten D, Hastie T, Tibshirani R: In: An introduction to statistical learning: with applications in R, 1st ed., New York, Springer, 2013, pp 2933

2.8 Manuscript 2: Statistical Presentation of Partial Surrogates

2.8.1 Abstract

The use of surrogate outcomes for inference has been well-studied in medical literature and remains controversial. On occasion however, it is necessary to develop models for prediction for which the true outcome of interest is infeasible to measure. The clinical motivation is perioperative acute kidney injury where serum creatinine change in the perioperative period is often used as a surrogate for kidney damage. This manuscript demonstrates the deficiency of serum creatinine change as a surrogate, due to its thresholded nature. These deficiencies provide the definition of a partial surrogate. Various statistical techniques for dealing with partial surrogates are examined and characterized, and practical guidance is provided for the analyst faced with using a partial surrogate outcome.

2.8.2 Introduction

Patient level clinical risk score development and associated decision support applications are vitally important to modern personalized medicine. For the majority of pathologies this process is straightforward. First, the disease process of interest is defined and data about relevant covariates are collected. This information is used to develop a statistical model which meets desired performance measures. However, when the disease process is difficult to directly measure, surrogate measurements are often used which prevent the use of simple risk score modeling methodology.

In 1989, Prentice defined necessary surrogate outcome criteria to ensure valid hypothesis testing [1]. Further work on surrogate outcome criteria has focused on the preservation of type I error rates for inference [2]. However, currently, surrogate outcome criteria for the development of risk scores remain undefined. These criteria will be developed in Section 2 of this work. After delineating criteria for the use of surrogate markers in risk score development, Section 3 will examine the increased modeling complexity associated with partial surrogacy situations. Partial surrogates are a class of markers that behave differently between patient subpopulations. In one subpopulation they may display a high association with the outcome of interest while in others they may display no association. This influences their ability to satisfy Prentices criteria for hypothesis testing. We will demonstrate that partial surrogate outcomes also complicate our proposed surrogate criteria for risk score prediction. Additionally, evaluating risk scores using a partial surrogate is complicated by the observation that the model which provides optimum discrimination for the surrogate outcome does not necessarily discriminate the true outcome well as will be demonstrated here. The implications of this observation for model selection and evaluation of likely clinical benefit will be described. Finally, section 4 of this manuscript will explore analytical challenges introduced by partial surrogacy theoretically and computationally.

In the course of this work, an analysis of perioperative acute kidney injury (AKI) will be performed to emphasize the clinical importance of our surrogate criteria for risk score modeling and to demonstrate the limitations and special considerations associated with partial surrogates in an applied analysis context.

2.8.3 Surrogate Outcomes in Risk Score Models

Clinical risk scores are commonly assessed in two ways. Firstly, by model discrimination, the degree to which a risk score is ordered similarly to the disease marker of interest. Secondly, by model calibration, a comparison of the magnitude of the risk score and the magnitude of the disease marker of interest. Risk scores that are well calibrated are simpler to implement and are traditionally considered ideal due to the observation that good calibration generally implies good discrimination. Unfortunately, risk scores built on surrogate outcomes rarely have good calibration. Typically, extensive knowledge of the relationship between the surrogate marker and the true outcome is necessary to facilitate post-hoc recalibration of the risk score in order to achieve acceptable calibration.

Now suppose that we are interested in developing a risk score, R, for a true clinical outcome, T, where R is any one-dimensional summary of a patients data that is intended to help quantify a patients disease state disposition. Next suppose that we are unable to measure T itself in the timeframe necessary to develop a useful decision support tool. Finally, suppose a surrogate outcome, S, is readily measurable and related to T either as a mediator or a consequence. What properties must S possess in order for a model based on S to result in a beneficial clinical risk score for T? While developing R, our goal will be to obtain a one-dimensional summary of the data that discriminates well, has good calibration, and maintaining some interpretability of the model coefficients as these are often used to generate hypotheses about potential mechanisms. A score, R, should provide higher scores for higher risk or more severely diseased patients uniformly over the entire range of plausible scores. We will refer to this last property as being clinically useful. Ideally, clinical utility should be consistent over the entire range of potential risk scores. Otherwise, Rs discriminatory ability might look favorable when examined over the entire population, despite R preforming poorly for a particular subset of patients. This could result in a net-benefit to the population at the expense of a particular group of individuals, raising questions about the ethical implementation of R for generalized patient care.

What criteria of S make the resulting risk score clinically useful? Following the pattern of Prentices first and second criteria for valid hypothesis testing [1], surrogate endpoints must display a relationship between the suspected risk factors to be included in the model, Z, and both the surrogate and true outcomes, S and T respectively. Stated more formally, the conditional distributions of S and T on Z must not be equal to the marginal distributions over Z.

P1. The proposed risk factor is related to the surrogate. $f(S|Z) \neq f(S)$

P2. The proposed risk factor is related to the true outcome. $f(T|Z) \neq f(T)$

The necessity of these two criteria, which when applied to risk score procedures will be referred to as R1 and R2 respectively, is fairly evident. A failure of criterion R1 suggests that the covariates included in the predictive model contain no information about the distribution of the surrogate. As such, models built on the surrogate would display little variation in the risk score R—Z and any variation observed would be random. A failure of criterion R2 suggests that the covariates are not related in any way to the distribution of the true outcome, and although $R|Z$ may display a rich variation, it would be expected that $f(TR,Z) = f(T)$.

The third Prentice criterion, stating that the conditional distribution of T on S needs to differ from the marginal distribution of T [1], is the basis for the development of our surrogate based risk score model criterion, R3. Formally stated,

P3. The surrogate measure is related to the true outcome. $f(T|S) \neq f(T)$.

However, for risk score development a more restrictive relationship between variables is necessary in order to obtain good discrimination and produce a clinically useful model. It is desirable that the distribution of $R|T$ be changing to favor more extreme values as T increases. Therefore, for some $T_1 < T_2$ corresponding to risk scores $R_1|T_1$ and $R_2|T_2$, we have that

R3. $P(R_1 < R_2|T_1,T_2) > 0.5$.

The R3 criterion promotes variation in R over different values of T. This ensures that, on average, the risk score is producing more extreme values when T is more extreme.

Ideally, the probability described in R3 would be large. Generally this occurs when the locational shift in the distribution of $R|T$ as T changes is large relative to its variance. Although not a strict requirement, having a risk score that is precise will naturally enhance its value. Another optional characteristic that enhances the utility of a model derives from Prentices fourth criterion [1]. The function of Prentices fourth criterion for hypothesis testing is to ensure that the surrogate captures the full effect of the covariate on the true outcome.

P4. The risk factors are related to the true outcome only through the surrogate. $f(TS,Z) = f(T|S)$

For risk score models this criterion is unnecessary as long as criterion R3 is satisfied.

However, the more information about the true outcome that is represented by the surrogate, the more likely it is that the surrogate will produce a risk score that is clinically useful.

In summary, our criteria for the development of a surrogate outcome based risk score are:

R1. The proposed risk factor is related to the surrogate. $f(S|Z) \neq f(S)$ R2. The proposed risk factor is related to the true outcome. $f(T|Z) \neq f(T)$ R3. The distribution of the risk score conditional on T needs to be shifting toward more extreme values amongst those at highest risk for disease $P(R_1 < R_2|T_1, T_2) > 0.5, T_1 < T_2$,

with P4 and the magnitude of the variance of $R|T$ relative to its distributional shift playing roles in determining the value of the resultant score.

These novel criteria encompass a surrogate outcomes minimum requirements to produce a valid risk score. In the next section we will begin to examine partial surrogates, and how the failure of some of these criteria in patient subpopulations can negatively impact risk score performance.

2.8.4 Theoretical Considerations Regarding Partial Surrogates

In the ideal situation, R1-R3 would hold in every subpopulation on which a risk score model is to be trained. In other words, it is beneficial if the phenotype defined by the relationship between Z, S, and T is homogenous throughout a population, **P**. If however there are subpopulations demonstrating differing phenotypes, extra care is required to maximize the benefit of risk score models and provide valid estimation procedures. When these heterogeneous subpopulations exist, we will redefine S to be a partial surrogate.

As an example, suppose you have collected data from **P** which is composed of two subpopulations V and I, defined by a latent indicator variable, l. In subpopulation V, R1-R3 hold, suggesting subpopulation **V** might produce a valuable risk score model. In subpopulation **I**, however, only R2 holds. This suggests that in subpopulation I, S is not meaningfully related to Z or T, and is therefore unlikely to result in a profitable risk score in this subpopulation.

The ideal method for risk score development when faced with a partial surrogate is not immediately apparent. One method is to use traditional modeling strategies in the full training dataset. In cases where the full dataset satisfies R1-R3 this approach is likely to result in valuable models. If l was known, an analyst might reasonably decide to use only the data from subpopulation V for model development, and then generalize the model to the entire population as appropriate. This second method relies on the relationship between T and Z being homogeneous over **P**. Homogeneity will occur if the subpopulations were defined completely at random. Alternatively, in cases where l is unknown, a latent variable mixture model can be used to produce a similar result. For the duration of this manuscript, l is assumed to be latent.

Given these two approaches, the analyst is forced to choose between the full-data approach and the mixture model approach. For inference and estimation, the choice is clear. Since failing to account for the partial nature of the surrogate will likely result in a violation of P4, the mixture model is preferable. For example consider a very simple partial surrogate where $T = S|V + \varepsilon_(T|S)$ and $S|I = \varepsilon_(S|I)$,

$$\varepsilon_i \sim N(0, \sigma_i)$$

and also $T|Z = \beta_(T|Z) + \beta_1 Z + \varepsilon_(T|Z)$ In this situation the surrogate is equal to truth plus error when a patient belongs to subpopulation V, but it is a random deviate when the patient is from subpopulation I. The relationship between T and Z is consistent across the entire population. Thus we have

$$E[T|S] = P(V)$$

$$S + P(I)$$

$$E[T] = P(V)S + (1 - P(V))E[T]$$

and also that $E[TS,Z] = P(V)S + P(I)E[TZ] = P(V)S + (1 - P(V))(\beta_(T|Z) + \beta_1 Z)$.

P4 requires that the distribution of T—S be the same as the distribution of T—S, Z, but even this simple partial surrogate violates that criterion as evidenced by the differing expectations.

However, for risk score modeling the decision is less clear. Using the full dataset and not accounting for the partial nature of the surrogate generally results in risk scores with lower variance due to higher effective sample size but higher bias due to the inclusion of training data from population I. The mixture model approach generally boasts reduced bias by correctly accounting for heterogeneous subpopulations but suffers higher variance due to diminished training set sample size. There are several aspects unique to a given partial surrogate situation that should affect the analysts decision regarding these modeling strategies.

When making the decision between using a traditional model or a mixture model, the first consideration is whether the added complexity of the mixture model approach is likely to be beneficial. The mixture models primary purpose is to estimate covariate/outcome relationships in the subpopulations separately. In order for this to practically improve the risk scores discrimination it needs to result in a different rank ordering of subjects compared to the traditional approach. This is likely to occur whenever the phenotype expressed in subpopulation I is substantively different than that in subpopulation V in terms of the relative magnitude of the associations between the covariates and outcome. This distinctness of subpopulation phenotypes simultaneously allows the expectation maximization (EM) algorithm used for model fitting to achieve adequate subpopulation separation while achieving a more appropriate ordering of predictions with respect to T.

The EM algorithm involves beginning with a prior probability of group assignment, fitting a model that is weighted by the prior probability to assess the likelihood of subpopulation membership, and calculating a posterior probability of subpopulation membership based on the prior and the likelihood. This is repeated until convergence is achieved with the posterior probabilities being used to generate the prior probabilities for the next iteration [3]. Substantial separation between subgroup phenotypes results in the EM algorithm calculating final posterior probabilities of subpopulation membership that are close to zero and one, suggesting there is good evidence in the data to direct each patients subpopulation assignment. When the subpopulations cannot be effectively separated, mixture model variance will be magnified, detracting from its utility and favoring the traditional modeling approach.

A second consideration affecting the development of partial surrogate based risk scores is how generalizable a subpopulation model based on **V** will be to the entire population **P**. If separation into subpopulations **I** and **V** is completely random, then any result obtained from subpopulation **V** should be fully generalizable. If subpopulations **I** and **V** are generated by a non-random process, however, neither modeling technique considered above is guaranteed to result in a clinically beneficial risk score, and additional external verification would be necessary to allow generalization.

The last major consideration that influences whether the mixture model approach is viable for risk score development with partial surrogates is the mixing proportion of the population. It is necessary to estimate what proportion of observations are from **V** versus the proportion from **I**. If the training data are composed almost entirely of data from **V**, the mixture model adds little benefit over the traditional model which ignores subpopulations. In contrast, if the data are almost entirely from **I**, there may not be enough information in the data to accurately fit a model for subpopulation V, which embodies the clinically relevant

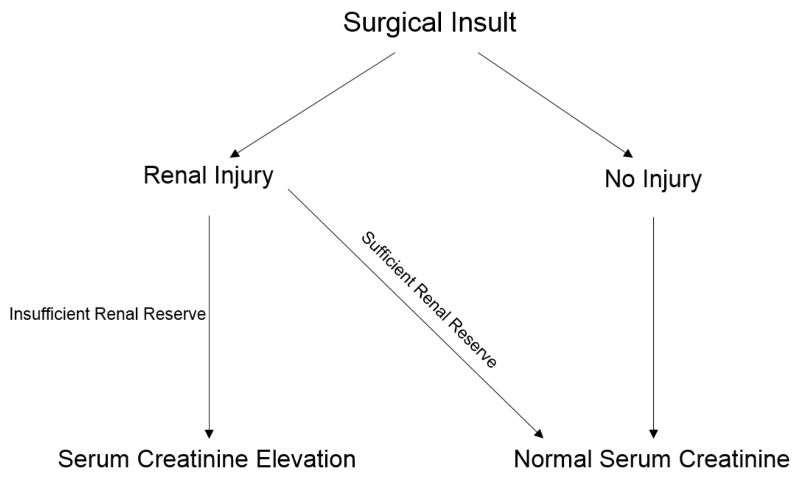


Figure 2.9: DAG displaying the proposed mechanism of partial surrogacy or serum creatinine as a marker of renal injury.

covariate/outcome relationship. In both of the situations described here partial surrogate based risk score models are unlikely to provide a benefit over the traditional modeling approach, because the available dataset does not contain enough information regarding the true relationship between covariates and the outcome of interest.

In summary there is no one-size-fits-all answer to dealing with partial surrogate outcomes. At times using the mixture model approach will be of great benefit. At other times the mixture approach is difficult or impossible to fit resulting in inferior performance when compared with the more traditional, non-mixture approach.

2.8.5 AKI Example of Risk Criteria

Biological Background

The goal of perioperative AKI research is to accurately assess the degree of kidney damage a patient suffers due to the physiologic stress of surgery. Perioperative AKI incidence rates range from 10 to 40% for major inpatient surgical procedures [4]. Perioperative AKI has been associated with increased short and long term mortality, increased hospital length of stay, increased risk of developing chronic kidney disease (CKD), and increased risk of developing dialysis dependence [4, 5]. Unlike other perioperative injuries, there is currently no direct biomarker of kidney injury or cell death. Therefore, perioperative acute kidney injury is diagnosed by comparing postoperative serum creatinine concentration to preoperative baseline serum creatinine concentration [4], figure 1. Despite repeated recent revisions to the diagnostic criteria for AKI [6-8] to increase sensitivity, there remains mounting evidence that patients sustain kidney damaged undetected by changes in serum creatinine, referred to as subclinical AKI [9]. This situation is illustrated well by the following example: it is not uncommon for a living kidney donor to experience little to no serum creatinine elevation despite removal of roughly 50% of their functional kidney mass [10]. Recent AKI biomarker studies demonstrated that subclinical AKI is also associated with an increased risk of dialysis and in-hospital mortality, suggesting it represents clinically significant levels of renal injury [11]. Creating a risk score which identifies patients suffering subclinical AKI in the immediate postoperative period would allow physicians to adjust these patients treatments to avoid nephrotoxic medications and optimize fluid status for kidney perfusion, possibly preventing morbidity and mortality.

The dramatic example of living donor kidney donation and minimal serum creatinine change demonstrates that healthy kidneys have the capacity to temporarily increase their filtration rate in times of physiologic stress, a characteristic known as renal functional reserve [12]. However, renal functional reserve is limited and can be exhausted [12]. In the subpopulation of patients, V, who overcome their renal functional reserve during perioperative episodes of kidney injury serum creatinine change would be detected, and associations between relevant risk factors and serum creatinine change would be strong, assuming all other serum creatinine modifying factors remain constant. In the subpopulation of patients, I, who do not overcome their renal functional reserve during episodes of kidney damage, only random or nonspecific changes in serum creatinine levels would be measured, and the associations between relevant AKI risk factors and serum creatinine change would be weak. With respect to the proposed risk score criteria outlined in section 2, this suggests that subpopulation V will largely satisfy P1-P4 and R1-R3, allowing for simultaneous estimation of associations and risk score generation. In contrast, subpopulation I will likely violate P1, P3, P4, R1 and R3, resulting in poor performance of risk indices based exclusively on this subgroup, biased coefficient estimates, and improper p-values from analyses based on the entire population.

If subpopulations I and V are defined based on exhaustion of renal functional reserve as we hypothesize, then it is important to recognize that likelihood of renal functional reserve exhaustion is not random. Young, healthy patients are less likely to overcome their substantial renal reserve than older patients with underlying disease [12, 13]. Therefore generalizing a risk score generated in subpopulation V to the entire population P requires validation of that score in the entire population. This validation can be accomplished by evaluating the partial surrogate based clinical risk scores discrimination of the true outcome, T. Although there is no gold standard marker for clinically significant kidney damage, one marker thought to be representative is the decline in kidney filtration rate at 90 days (eGFR90) [14, 15]. The 90 days between surgery and the time this GFR is calculated allows the kidneys to recover from acute injury if possible and reestablish an equilibrium serum creatinine concentration. Indeed, current clinical guidelines recommend that patients who experience AKI should routinely have 90 day eGFR evaluation to assess recovery verses progression to permanent kidney damage [8]. Therefore, in this analysis eGFR90 will be considered the true outcome, T.

Dataset Description Data and Models: The data used in this analysis are from 4737 patients who underwent cardiac surgery at a large academic medical center from November 2009 through June 2015. Institutional IRB approval was obtained prior to performance of all analyses. In this dataset, 1268 patients had 90±15 day eGFR90 measurements available.

Ten preoperative and intraoperative traits were selected a priori for inclusion in the analysis including age, body mass index, a diagnosis of diabetes, baseline kidney (glomerular) filtration rate, baseline hemoglobin concentration, volume of intraoperative urine output,

	Estimates Subpopulation I	Estimates Subpopulation V
BMI	0.047 (0.038,0.057)	0.25 (0.206,0.295)
Total Urine Output	-0.041 (-0.052,-0.03)	-0.147 (-0.193,-0.1)
Total Fluids Given	-0.015 (-0.025,-0.004)	-0.134 (-0.179,-0.09)
Age	0.063 (0.051,0.074)	0.213 (0.168,0.258)
Baseline eGFR	0.11 (0.097,0.122)	0.388 (0.344,0.433)
Hemoglobin	-0.063 (-0.074,-0.053)	-0.253 (-0.297,-0.209)
Max Intraoperative Lactate	0.06 (0.048,0.073)	0.206 (0.161,0.25)
Diabetes	-0.002 (-0.013,0.008)	-0.109 (-0.152,-0.066)
Length of Surgery	0.072 (0.058,0.085)	0.193 (0.144,0.241)
Emergency Surgery	0 (-0.011,0.011)	0.05 (0.012,0.087)

Table 2.11: Coefficients resulting from the application of the mixture model to the perioperative AKI dataset. The column on the right represents the coefficients from subgroup V and are noticeably larger in absolute magnitude than those on the left.

volume of intraoperative intravenous fluid administered, maximum measured intraoperative plasma lactate level, length of surgery, and an indicator for emergent surgery. These variables were chosen as well-established predictors of AKI and therefore were considered likely to be valuable predictors of serum creatinine change from baseline [16-20].

For the purposes of model comparisons, a linear model and a two-component mixture of linear models were fit. The residual error of the two mixture components was not constrained. The linear model risk score is the models predictions. For the mixture model, the risk score is the prediction from the single component of the mixture that is post-hoc identified to be associated with subpopulation V. Each model was evaluated based on the following metrics: the AUC for a true outcome greater than 20 and the Spearmans correlation. These metrics represent frequently used methods for model assessment in clinical literature. The first metric is a common method of risk score implementation and is based on the presumtion that a change of 20 in eGFR90 is clinically meaningful. The second metric measures discrimination without requiring an arbitrary cutoff.

The mixture model resulted in moderately well differentiated clusters, a relative entropy=0.607, 728 patients being modally assigned to the V subpopulation and 4009 patients being assigned to the I subpopulation. The linear model found all the factors to be significant except for history of diabetes and emergency surgery, which it found to be marginally significant (p=0.085 and 0.063 respectively). The mixture component found all the risk factors to be significant with the exception of emergency surgery (p=0.072). However, the magnitude of the coefficients for the linear model were attenuated by an average of 42.8% (range=[19.2%, 68.1%]), which is consistent with subpopulation Is phenotype being null or attenuated relative to subpopulation V. The model coefficients are given in Table 1. In addition, the mixture model represented a substantial improvement in fit over the linear model with BICs of 2131.3 and 5034.3 respectively.

The area under the ROC curve was calculated for each of the candidate risk scores and for the gold standard of the observed serum creatinine change for the prediction of an eGFR90 decline greater than 20 mL/min/1.73 m2. The observed serum creatinine change had the worst estimated AUC of 0.608 (0.572, 0.645), although not significantly worse than that of the linear model 0.633 (0.595, 0.672), p=0.262. The mixture model component yielded the best AUC of 0.678 (0.641, 0.715), which was a significant improvement over both the observed creatinine change and the linear model, p=0.002 and p <0.001 respectively. In addition, the ROCs were calculated for each candidate risk score for the prediction of a serum creatinine increase greater than 0.3 mg/dL, a common clinical cutoff for AKI [6, 8]. The result was an AUC of 0.602 (0.582, 0.623) for the mixture and 0.663 (0.644, 0.682) for the linear model, p <0.001. The ROCs for both endpoints are given in Figure 2.

The improvement due to using the mixture components prediction as a risk score for eGFR90 is further demonstrated by looking at Spearmans rank correlation. The correlation between the observed serum creatinine change and the observed eGFR90 change was 0.231 (0.204, 0.258). For the linear model the correlation was 0.223 (0.196, 0.250). For the mixture component the correlation was 0.305 (0.280, 0.331). These values were compared via a permutation test showing a significant improvement by the mixture model over the observed value and the linear models prediction, p=0.035 and 0.020 respectively. The low

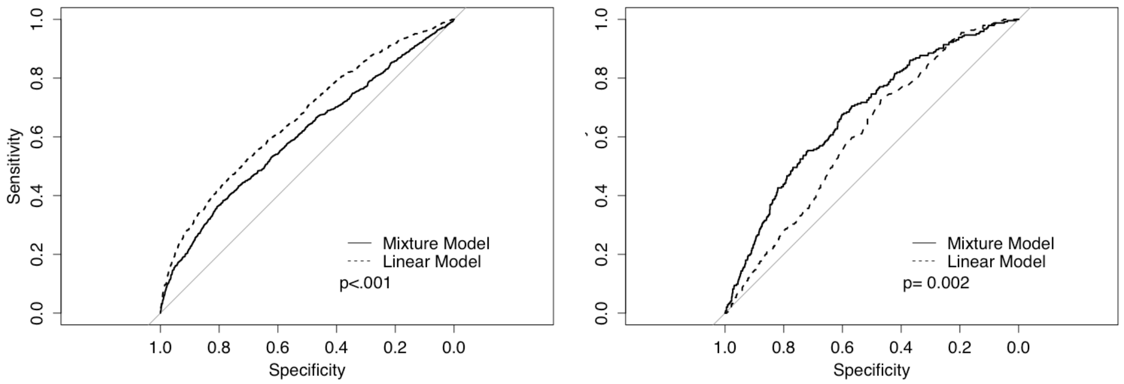


Figure 2.10: ROC curves for the linear and mixture model for the prediction of a perioperative change in serum creatinine concentration >0.03 mg/dL and an eGFR decrease of greater than 20 mL/min/1.73m2 at 90 days postoperatively. Note how using the flawed, partial surrogate for validation would lead to improper model choice.

values of these correlations is due to the fact that the majority of surgical patients sustain no kidney injury and thus any change in their eGFR is truly random, i.e. only a small portion of the populations eGFR changes are ordered by something other than random chance, so despite the low correlation the improvement provided by utilizing the partial surrogate is substantial.

This analysis demonstrates a major issue in the development of risk scores usingpartial surrogate outcomes. If the partial surrogate nature of serum creatinine change had gone unrecognized in this analysis, the analyst would likely look to how well various models discriminate with respect to serum creatinine change as a preferred method for both model selection and characterization. The ROC analysis demonstrates that the analyst would then conclude that the linear model was clearly superior to the mixture model because its risk score is ordered more similarly to the surrogate measure. However, the ROC of the true outcome, eGFR90, shows the true relationship is reversed and that the mixture model produced a superior ordering. It is critical to identify partial surrogates and account for them appropriately, since there would be no indication of this flaw in analysis if model performance was judged solely on its ability to predict the surrogate outcome, postoperative serum creatinine elevation.

Simulation Studies

To further explore how the relationships between the covariates and the surrogate influence the benefit associated with a mixture modeling approach compared with traditional linear modeling, the perioperative AKI clinical data previously described was used to generate two simulation studies. For each simulation, a two-component mixture of linear models and a traditional linear model were fit. As before, the residual error of the two mixture components was not constrained. The linear model risk score was the models predictions. The mixture models risk score was the prediction from the single component of the mixture that was posthoc identified to be associated with subpopulation **V**. The partial surrogate, maximum 48-hour postoperative change in serum creatinine concentration compared to preoperative baseline serum creatinine concentration, was represented by S. Each simulation used the fitted models on the perioperative AKI dataset as the data generating mechanism for simulation, figure 3. New outcomes were generated for each simulation using standard errors and coefficient values estimated from the observed data. In each simulation, each participants posterior probability of subgroup membership was taken from the fitted clinical data and used to generate a new group assignment via a Bernoulli draw. Surrogate outcomes for those assigned to group **V** were drawn from the normal distribution suggested by the component of the fitted clinical mixture model representing group **V**. If assigned to group **I**, the surrogate was drawn differently in each simulation. The true outcome was also drawn from the fitted model representing group **V** and then normalized to have mean 0 and standard deviation of 50 in order to make the values similar to what is seen with eGFR90. These simulations are represented graphically in Figure 3. Each simulation was repeated 1000 times and the resultant linear and mixture models were compared using the metrics discussed above: AUC of predicting a decrease in the true outcome $>$ 20 units, Spearmans correlation between the risk score and the true outcome, and an additional metric, percent mean square error(MSE) reduction of the coefficients from the model generating coefficients.

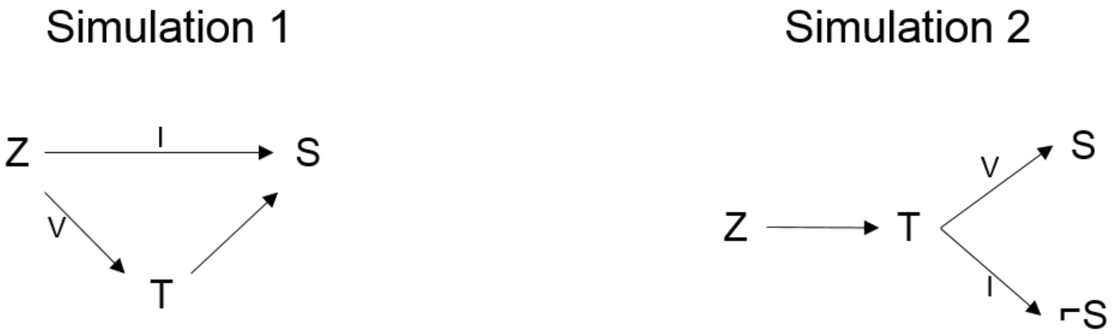


Figure 2.11: Simulation 1 mimics a situation where there are two mechanistic pathways between the covariates and the outcome. Simulation 2 is a situation in which the surrogate is a resultant of the true outcome of interest on one pathway and is unrelated to it in the other, similar to our proposed AKI mechanism.

Our first simulation exemplifies a scenario where there are substantial differences in the coefficient magnitudes between the two subgroups. Covariate coefficients for the model representing subpopulation **I** were randomly drawn from a normal distribution with mean zero and standard deviation equal to of the range of coefficient values observed in the clinical data (0.176). The intercept was fixed at the fitted value of the mixture component representing group **I**.

$$\begin{aligned}\beta_{R,i} \sim N(0,.176), i = 1,..,10\\ T|Z \sim N(Z\hat{\beta}_V, \hat{\sigma}_V)\\ S|Z,V \sim N(Z\hat{\beta}_V, \hat{\sigma}_V)\\ S|Z,I \sim N(Z[\beta_{I,1}, \beta_{R,1}, \ldots, \beta_{R,10}], \hat{\sigma}_I)\end{aligned}$$

The results for each model in the simulation are given as mean (0.05 quantile, 0.95 quantile). The AUCs of the linear model for discriminating a true outcome greater than 20 encompassed almost the entire range of potential values (0.737 (0.518, 0.923)). For the mixture approach, the AUCs were substantially more consistent (0.950 (0.908, 0.982)). This result represents an AUC difference of 0.213 (0.034, 0.433). The Spearmans rank correlation for the linear model displayed a wide range of values from negative correlation up to more reasonable values for the particular problem (0.152 (-0.091, 0.349)). Again the mixture model approach yielded results that were more consistent and clinically useful (0.501 (0.471, 0.526)). This represents a difference in rank correlation of 0.349 (0.146, 0.603). Lastly, with respect to the MSE of estimated coefficients, the linear and mixture model results were 0.048 (0.024, 0.077) and 0.004 (0.002, 0.008) respectively. This represents a relative reduction in the MSE of the coefficient estimates of 89.1% (74.3%, 97.0%) and an absolute reduction of 0.043 (0.019, 0.072).

The second simulation represents a scenario where no relationship between the covariates and the surrogate outcome exists in subpopulation **I**. This scenario is analogous to our clinical example when the body compensates for kidney damage using renal functional reserve, producing a subpopulation of patients, **I**, that suffer kidney damage but have no change in serum creatinine levels. Physiological compensatory mechanisms such as renal functional reserve can result in no relationship between Z and S in subpopulation **I**. In these situations, the surrogate outcomes in **I** are simply normal deviates about zero.

$$
\begin{aligned}
T|Z &\sim N(Z\hat{\beta}_V, \hat{\sigma}_V) \\
S|Z,V &\sim N(Z\hat{\beta}_V, \hat{\sigma}_V) \\
S|Z,I &\sim N(0, \hat{\sigma}_I)
\end{aligned}
$$

For this simulation, the AUCs of the linear and mixture models were 0.928 (0.873, 0.972) and 0.949 (0.906, 0.982) respectively, resulting in a difference of 0.021 (0.002, 0.044) between the mixture and the linear models. The Spearmans rank correlations were 0.394 (0.333, 0.451) and 0.492 (0.454, 0.524) respectively, producing a difference of 0.099 (0.065, 0.133). The MSE of the coefficients were 0.031 (0.028, 0.034) and 0.005 (0.002, 0.009) respectively. This represents a reduction in the relative MSE of the coefficients by using the mixture model method of 83.0% (71.5%, 92.1%), and an absolute reduction of 0.026 (0.022, 0.029).

This second simulation demonstrates that the mixture models gain in discriminatory ability is mildly decreased when the surrogate outcomes from group **I** are random with respect to T. However, even in this scenario, the mixture model displays a substantial reduction in MSE of the coefficient estimates when compared to the linear model.

Discussion

Whether the goal of an analysis is inference or prediction of a risk model, the value of recognizing partial surrogacy of an outcome marker is clear. In our clinical example of perioperative acute kidney injury it was demonstrated that treating serum creatinine change as a full surrogate rather than a partial one led to the erroneous conclusion that the linear model approach was much better than a mixture model at measuring kidney injury, represented by eGFR90. Change in serum creatinine from baseline when used as a surrogate for true kidney damage. Prior to this work, no account has been given to the partial surrogate nature of serum creatinine change, which has resulted in biased estimates of covariate effects resulting from a serious failure of the Prentice criteria.

The simulation studies included here are meant to highlight the complexity of the decision on how to model a partial surrogate for the development of a risk score This decision is heavily influenced by a mixture models ability to resolve subgroup **V** from subgroup **I** and by the relationships between covariates and the surrogate outcome in subpopulation **I**. In practice, the only way an analyst can quantify these issues is by fitting a mixture model whenever partial surrogacy is suspected. By inspecting the fitted mixture model, the analyst will then be able to assess the models entropy and the clinical significance of the difference between the phenotypes estimated by the mixture model. This provides the analyst with a better understanding of the effect partial surrogacy has on her potential risk score model. Ultimately, the mixture model approach to a suspected partial surrogate provides valuable insight into whether Prentices criteria and the criteria proposed here are likely to be satisfied.

References

1. Prentice, R.L., Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med, 1989. 8(4): p. 431-40.

2. Weir, C.J. and R.J. Walley, Statistical evaluation of biomarkers as surrogate endpoints: a literature review. Stat Med, 2006. 25(2): p. 183-203.

3. Dempster AP, L.N., Rubin DB, Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B, 1977. 39(1): p. 1-38.

4. Calvert, S. and A. Shaw, Perioperative acute kidney injury. Perioper Med (Lond), 2012. 1: p. 6.

5. Thakar, C.V., Perioperative acute kidney injury. Adv Chronic Kidney Dis, 2013. 20(1): p. 67-75.

6. Bellomo, R., et al., Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. Crit Care, 2004. 8(4): p. R204-12.

7. Mehta, R.L., et al., Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care, 2007. 11(2): p. R31.

8. Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group, KDIGO clinical practice guideline for acute kidney injury. Kidney Int Suppl, 2012. 2: p. 1-138.

9. Ronco, C., J.A. Kellum, and M. Haase, Subclinical AKI is still AKI. Crit Care, 2012. 16(3): p. 313.

10. Najarian, J.S., et al., 20 years or more of follow-up of living kidney donors. Lancet, 1992. 340(8823): p. 807-10.

11. Haase, M., et al., The outcome of neutrophil gelatinase-associated lipocalin-positive

subclinical acute kidney injury: a multicenter pooled analysis of prospective studies. J Am Coll Cardiol, 2011. 57(17): p. 1752-61.

12. Sharma, A., M.J. Mucino, and C. Ronco, Renal functional reserve and renal recovery after acute kidney injury. Nephron Clin Pract, 2014. 127(1-4): p. 94-100.

13. Fliser, D., et al., Renal functional reserve in healthy elderly subjects. J Am Soc Nephrol, 1993. 3(7): p. 1371-7.

14. Coca, S.G., Is it AKI or nonrecovery of renal function that is important for long-term outcomes? Clin J Am Soc Nephrol, 2013. 8(2): p. 173-6.

15. Ishani, A., et al., The magnitude of acute serum creatinine increase after cardiac surgery and the risk of chronic kidney disease, progression of kidney disease, and death. Arch Intern Med, 2011. 171(3): p. 226-33.

16. Berg, K.S., et al., How can we best predict acute kidney injury following cardiac surgery?: a prospective observational study. Eur J Anaesthesiol, 2013. 30(11): p. 704-12.

17. Billings, F.T.t., et al., Obesity and oxidative stress predict AKI after cardiac surgery. J Am Soc Nephrol, 2012. 23(7): p. 1221-8.

18. Huen, S.C. and C.R. Parikh, Predicting acute kidney injury after cardiac surgery: a systematic review. Ann Thorac Surg, 2012. 93(1): p. 337-47.

19. Kim, W.H., et al., Simplified clinical risk score to predict acute kidney injury after aortic surgery. J Cardiothorac Vasc Anesth, 2013. 27(6): p. 1158-66.

20. Parolari, A., et al., Risk factors for perioperative acute kidney injury after adult cardiac surgery: role of perioperative management. Ann Thorac Surg, 2012. 93(2): p. 584-91.

Chapter 3

The evaluation of multiple therapies: Challenging the status quo in clinical trial design

3.1 Introduction to the Chapter

As the FDA broadens the classes of clinical trial design that it is willing to admit. A lot of research is taking place into how to do clinical trials better in the world of drug development. Unfortunately, even in the world of novel drug discovery these developments have been slow to catch on. Outside of that high-dollar, research intensive arena the field of clinical trials is largely still composed of tried and true frequentist, non-adaptive designs. These trials largely focus on how best to apply existing therapies to optimize some clinical outcome.

Tuning an intervention to achieve maximum benefit is not an easy task. If you consult the experts on any given intervention, you will often find vast differences in opinions on how it is best implemented. Despite differing opinions on how an intervention is best applied, the standard trial designs rarely incorporate more than one or two intervention groups. In this chapter we examine how the incorporation of many arms in a trial can increase efficiency and give substantial amounts of additional information about what the optimal intervention truly is. This approach is an extension of the platform-type trials recently advocated by Berry etal.

In our application of platform type trials, a novel adaptive randomization scheme is implemented. This method revolves around using the likelihood ratio from an empirically estimated density of treatment effects to set the randomization probabilities. In addition, recognition of the heterogenous nature of treatment effects when there are a variety of potential implementation protocols requires revisitation of the traditional type I error rate and power framework in which clinical trials are done since it is likely that any treatment effect is completely null. As such we investigate the use of Type S error rate control suggested by Gelman et.al and a corresponding quantity we deem “fuzzy power.”

Lastly, The simulation studies justifying this work have been incorporated into a Shiny app that allows the user to specify virtually every parameter of the simulation. This tool could be used as a template for the clinical trial design phase allowing user-specified treatment effect distributions, run-in period size, total sample size and others.

The primary clinical impetus for this work is the analysis of the efficacy of perioperative beta blockade for the prevention of perioperative cardiac ischemia. In this work a historical overview of the clinical development of the problem is undertaken. It is demonstrated how the affinity for static, two-arm designs lead to evidence that is ultimately inconclusive as to the intervention’s safety. Included in this work is a meta-analysis of RCTs that examine perioperative beta blockade any demonstrate the effect of improper pooling over heterogenous treatment effects ultimately leading to a conclusion that is entirely sensitive to the chosen statistical model.

3.2 Analysis of Perioperative Beta Blockade

The following manuscript was written to illustrate the problems inherent with using large two-arm studies as the final word in scientific evidence, especially in the face of significant heterogeneity. For the reader unfimiliar with the history of perioperative beta blockade, perioperative beta blockade is a proposed prophylactic treatment meant to reduce the incidence and severity of perioperative cardiac ischemia. Cardiac ischemia occurs when there is a supply and demand mismatch in the amount of oxygen delivered to the heart muscle (i.e. the heart requires more oxygen than is being supplied by its blood supply). The circumstances behind supply shortfalls are generally difficult to modify. For example, if someone has excessive bleeding during surgery so that their blood pressure falls and their heart muscle is not adequately profusing with blood, we do everything possible to restore adequate oxygen supply however these incidental shortfalls are often unavoidable. The promise of beta blockade is that it functions on the demand side of the equation. By slowing the patient's heart rate, the cardiac muscle requires less oxygen to maintain normal function.

Early observatinal studies and several small clinical trials showed promise. But a large confirmatory study, POISE, called into question the treatment's safety implying that it lead to a higher rate of ischemic stroke related to decreased vascular pressure. Many practitioners dismissed the results of the POISE trial because they disagreed with the relatively large amount of beta-blocker given as part of its protocol. Since POISE virtually no new studies have been conducted on beta-blockade as IRB approvals are hard to get. Meta-analyses disagree on what the current state of evidence is as the conclusion is based entirely on the rigidity of the chosen statistical model as shown in the following.

3.3 Manuscript: Metoprolol versus other β-Blocking Agents in Perioperative β-Blockade

3.3.1 Abstract

Background: Recent observational studies suggest that the association between perioperative β-blockade and increased risk of mortality and stroke varies based on β-blocker utilized. Metoprolol, the β-blocker utilized in the POISE trial, is associated with the highest risk of perioperative adverse events. No previously published meta-analysis has accounted for this heterogeneity of treatment effect.

Methods: Two meta-analyses of randomized controlled trials were performed examining initiation of perioperative β-blockade stratified by medication. Outcomes of interest included non-fatal stroke, non-fatal myocardial infarction, short term (30 days) and long term ($>$ 6 months) mortality.

Results: When short term outcomes are examined, metoprolol, but not other β-blockers, is associated with a statistically significantly decreased risk of myocardial infarct (p = 0.001), and increased risk of non-fatal stroke (p = 0.037) and short term mortality (p =

0.036). Support for these associations is almost completely derived from POISE trial data. Long term outcomes demonstrate a statistically significant difference in the effect of metoprolol versus that of other β-blockers (p = 0.049). A protective effect in long term mortality (p = 0.034) was found for β-blockers other than metoprolol.

Conclusions: The effect of perioperative β-blockade initiation varies by medication, restricting the generalizability of previous meta-analyses. In addition, previously utilized 30 day endpoints may fail to capture the complexity of postoperative mortality. The current state of evidence suggests that treatment with β-blockers other than metoprolol may have a protective effect on long term mortality.

3.3.2 Introduction

For years there has been a controversy surrounding initiation of β-blockade prior to non-cardiac surgery. Multiple randomized controlled trials have found a decreased risk of myocardial infarction (MI) with perioperative β-blocker initiation.[1, 2] However, the POISE trial also found an increased risk of perioperative stroke and 30 day mortality with treatment. Recently, the discovery of apparent scientific misconduct has resulted in nullification of DECREASE trial data, increasing the controversy and confusion regarding possible risks and benefits of perioperative β-blockade initiation. Additionally, since the publication of POISE, three large observational studies involving more than 140,000 patients have suggested that the association between perioperative β-blockade and increased adverse event rates varies based on β-blocker utilized.[3-5] These studies consistently find metoprolol more strongly associated with adverse events than other β-blockers. Given these new findings, reanalysis of the remaining randomized controlled trial evidence, stratified by medication, is needed.

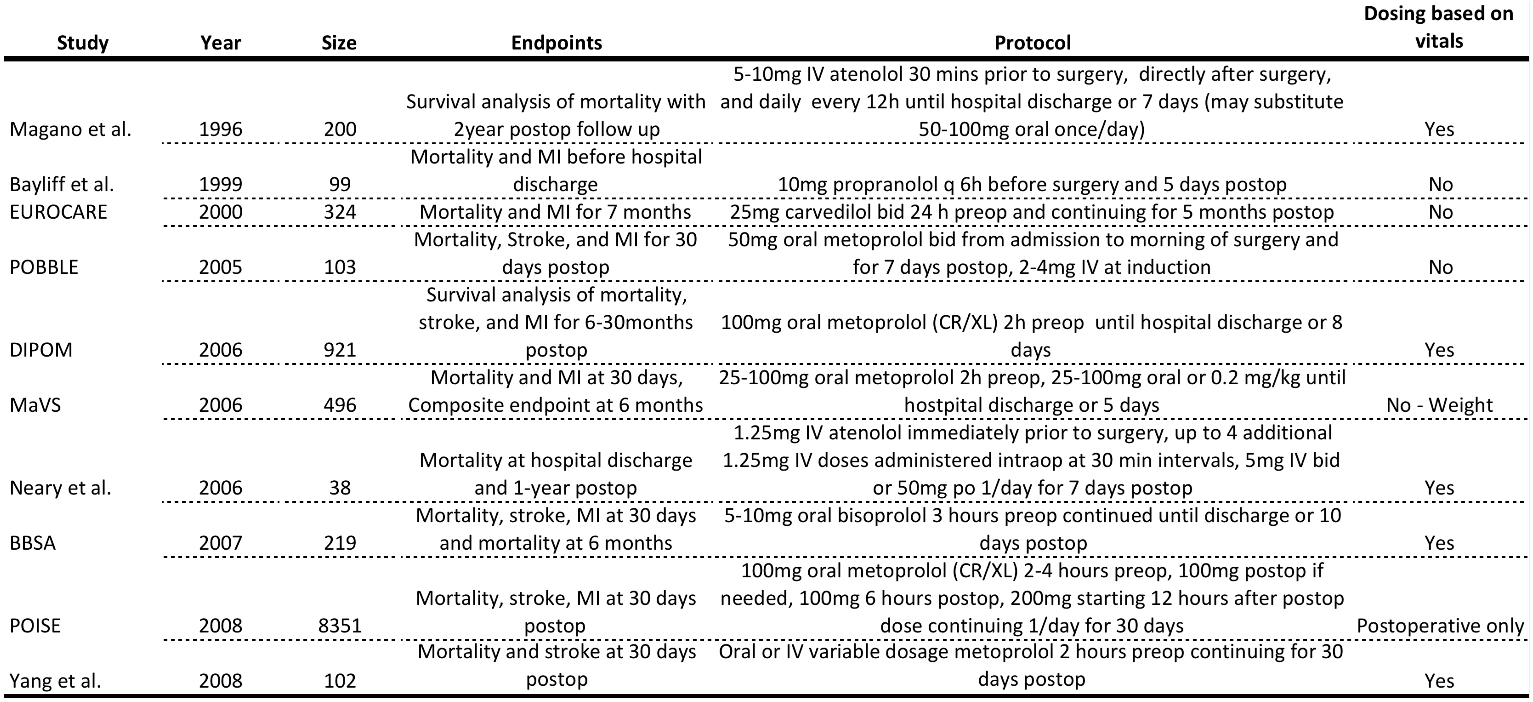
Patients who develop perioperative ischemia may be at increased mortality risk for years.[6, 7] If perioperative β-blocker treatment is protective with respect to myocardial ischemia, as some studies have suggested,[1, 8] then a 30 day follow up period is unlikely to capture the entire association of β-blocker treatment on mortality. It is possible that initiation of perioperative β-blockade results in increased short term mortality secondary to perioperative stroke but provides decreased long term mortality secondary to protection from myocardial ischemia. Failing to examine long term mortality endpoints prevents detection of such a crossing-hazards phenomenon.

A recent meta-analysis by Bouri et al.[9] examining initiation of perioperative β-blockade and the risk of adverse events after non-cardiac surgery has generated significant discussion, including a front page report in Anesthesiology News, as well as articles in the British Medical Journal and Heartwire.[10-12] There are three major limitations of the Bouri et al. study. First, the study is not stratified by medication. It therefore assumes all β-blockers have similar risk profiles, which is contradictory to available observational data.[3-5] Secondly, its mortality evaluation is limited to the 30 day post-operative period, preventing the detection of any long term mortality effects associated with treatment.[9] Finally, 92% of the weight in the Bouri et al. analysis was from studies utilizing metoprolol (calculation not shown), but the conclusions drawn are generalized to all β-blockers.

In this work we present meta-analyses examining the effect of perioperative initiation of metoprolol versus that of other β-blockers on perioperative non-fatal MI, perioperative non-fatal stroke, short term (30 days) and long term ($>$ 6 months) mortality. These results have a more clinically meaningful interpretation than previous unstratified analyses given the mounting evidence that metoprolol has an especially poor risk profile when compared to other β-blockers. They also provide a more complete picture of the association between perioperative β-blockade and long term mortality than previous analyses.

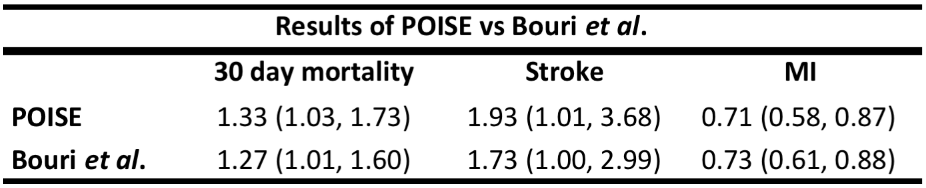
3.3.3 Materials and Methods

Literature searches were performed on PubMed and Google Scholar to identify all trials comparing β-blocker treatment to no treatment or placebo. Hand searches of previous meta-analyses were also performed. Trials were excluded from the short term endpoint



Study	Year	Size	Endpoints	Protocol	Dosing based on vitals
Magano et al.	1996	200	Survival analysis of mortality with 2year postop follow up	5-10mg IV atenolol 30 mins prior to surgery, directly after surgery, and daily every 12h until hospital discharge or 7 days (may substitute 50-100mg oral once/day)	Yes
Bayliff et al.	1999	99	Mortality and MI before hospital discharge	10mg propranolol q 6h before surgery and 5 days postop	No
EUROCARE	2000	324	Mortality and MI for 7 months	25mg carvedilol bid 24 h preop and continuing for 5 months postop	No
POBBLE	2005	103	Mortality, Stroke, and MI for 30 days postop	50mg oral metoprolol bid from admission to morning of surgery and for 7 days postop, 2-4mg IV at induction	No
DIPOM	2006	921	Survival analysis of mortality, stroke, and MI for 6-30months postop	100mg oral metoprolol (CR/XL) 2h preop until hospital discharge or 8 days	Yes
MaVS	2006	496	Mortality and MI at 30 days, Composite endpoint at 6 months	25-100mg oral metoprolol 2h preop, 25-100mg oral or 0.2 mg/kg until hostpital discharge or 5 days	No - Weight
Neary et al.	2006	38	Mortality at hospital discharge and 1-year postop	1.25mg IV atenolol immediately prior to surgery, up to 4 additional 1.25mg IV doses administered intraop at 30 min intervals, 5mg IV bid or 50mg po 1/day for 7 days postop	Yes
BBSA	2007	219	Mortality, stroke, MI at 30 days and mortality at 6 months	5-10mg oral bisoprolol 3 hours preop continued until discharge or 10 days postop	Yes
POISE	2008	8351	Mortality, stroke, MI at 30 days postop	100mg oral metoprolol (CR/XL) 2-4 hours preop, 100mg postop if needed, 100mg 6 hours postop, 200mg starting 12 hours after postop dose continuing 1/day for 30 days	Postoperative only
Yang et al.	2008	102	Mortality and stroke at 30 days postop	Oral or IV variable dosage metoprolol 2 hours preop continuing for 30 days postop	Yes

Table 3.1: Studies Included in Meta-Analyses and their endpoints



Results of POISE vs Bouri *et al*.			
	30 day mortality	**Stroke**	**MI**
POISE	1.33 (1.03, 1.73)	1.93 (1.01, 3.68)	0.71 (0.58, 0.87)
Bouri *et al*.	1.27 (1.01, 1.60)	1.73 (1.00, 2.99)	0.73 (0.61, 0.88)

Table 3.2: Comparison of POISE versus Bouri et al. results

analysis if:

- β-blocker treatment did not begin preoperatively or did not extend into postoperative period
- The study examined patients receiving high-risk cardiac procedures
- The publication was not available in English

All included studies were randomized, blinded, and based on intention-to-treat analysis. Additional data was obtained directly from authors of several studies. Table 1 summarizes identified studies.[1, 7, 13-19] To capture additional data on long term mortality, a second literature review was conducted to identify studies that reported mortality outcomes 6 months post-surgery. The search was performed in a fashion similar to that previously described. The previously outlined exclusion criteria were used. Identified long term mortality studies are also presented in Table 1.[7, 14, 16, 18, 19] Since these studies do not have equal follow-up periods, mortality data were taken from each studys endpoint to maximize power and minimize the risk of missing crossing-hazards. Results from this part of the analysis can thus be interpreted as mortality at some average follow-up time ¿ 6 months.

Due to the binary nature of each outcome considered, we were able to reconstruct individual patient level predictor data to use in our analyses as compared to traditional meta-analyses which rely on published summary statistics. Both meta-analyses were completed using stratified mixed-effects logistic regression modeling with random intercepts to account for differing baseline event rates among studies. The advantage of reconstructing individual patient data and utilizing a stratified mixed-effects model is that this method prevents the averaging of effect sizes across groups known to be different, which occurs in a traditional meta-analysis, while providing a better estimate of between-subject variability. In addition, it removes heterogeneity of treatment effect due to metoprolols potentially unfavorable risk profile and allows for partial sharing of control data across study groups. Although a model that included random intercepts was considered, there was not sufficient

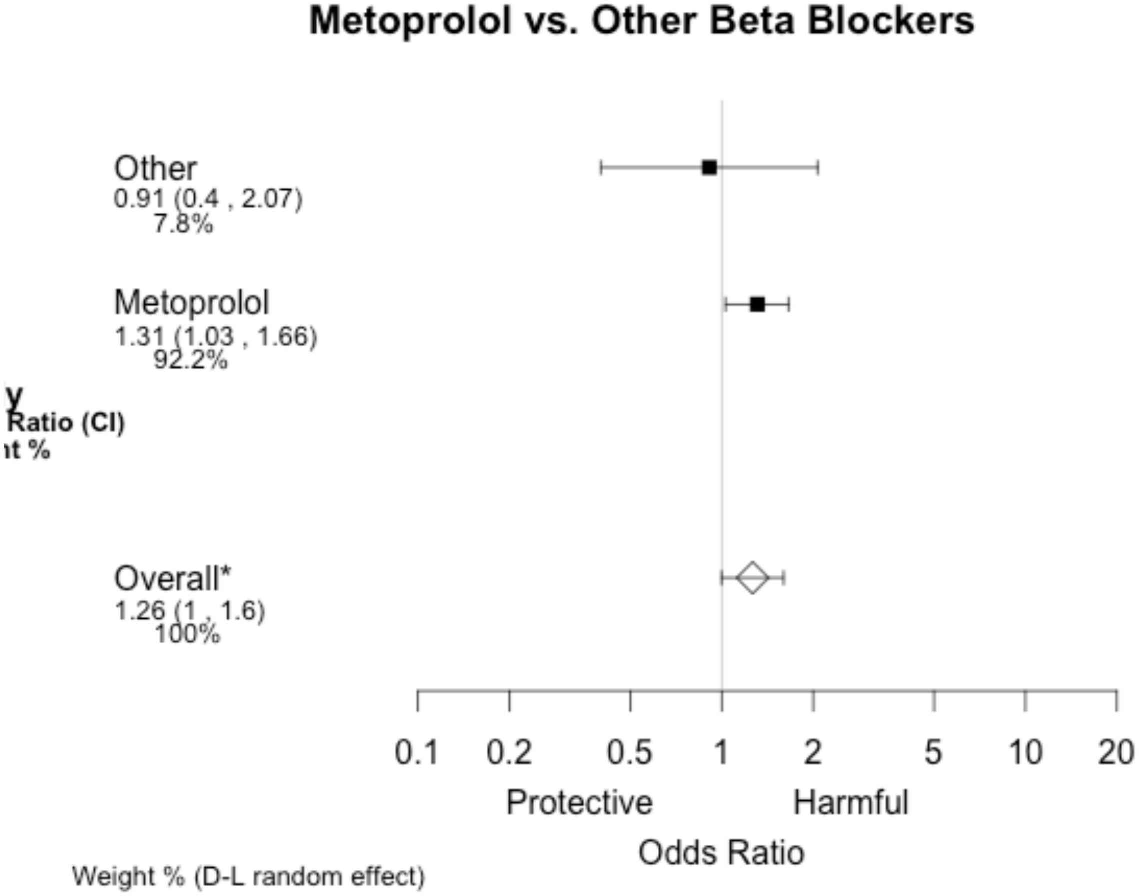
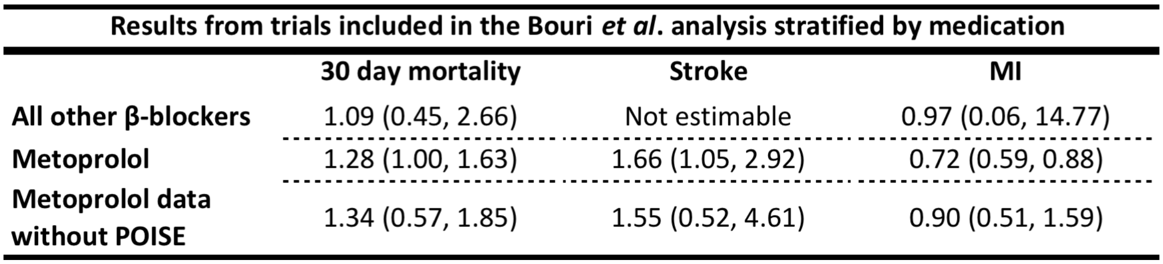


Figure 3.1: Bouri et al. analysis grouped by trial medication. Note the high weight ascribed to metoprolol studies along with the similarity of the metoprolol and overall intervals.

data to warrant the additional increase in model complexity. The results of these analyses are odds ratios but, owing to the rarity of the conditions under study, they are excellent approximations to the relative risk.[20]

The inter-study variation (ρ) estimated for the short term outcomes regression models were 28.1%, 9.3%, and 29.9% for mortality, stroke, and MI respectively. These values are considered mild.[21] for the long term mortality regression model was estimated to be 44.3%. Using as a measure of heterogeneity is preferable to the usual I2 statistic in this case due to the low event rates in these studies.[22] All statistical analyses were performed in Stata version 12.1 (StataCorp, College Station, TX).



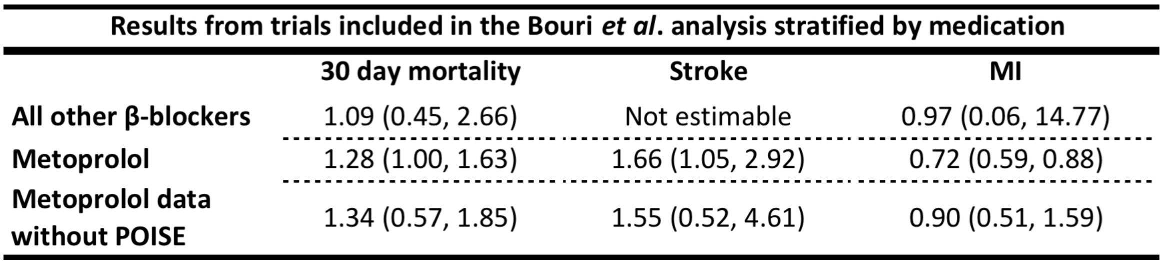
Results from trials included in the Bouri *et al*. analysis stratified by medication			
	30 day mortality	**Stroke**	**MI**
All other β-blockers	1.09 (0.45, 2.66)	Not estimable	0.97 (0.06, 14.77)
Metoprolol	1.28 (1.00, 1.63)	1.66 (1.05, 2.92)	0.72 (0.59, 0.88)
Metoprolol data without POISE	1.34 (0.57, 1.85)	1.55 (0.52, 4.61)	0.90 (0.51, 1.59)

Table 3.3: Results of stratified meta-regression analysis for 30 day endpoints. Note the similarity of metoprolol group to POISE results as well as the complete lack of evidence when POISE is excluded. Also note there is virtually no evidence in the Other beta-blocker group.

3.3.4 Results

To assess the current state of evidence regarding benefits and harm of perioperative initiation of β-blockade, a stratified meta-analysis of randomized controlled trial data was performed. The short term outcome results are presented in Figure 1A-C. The metoprolol outcomes show a statistically significant decrease in perioperative MI (p = 0.001), and a significant increase in perioperative stroke (p = 0.037) and short term mortality (p = 0.036) with treatment. These results bear a striking similarity to the following outcomes reported by POISE: relative risk of non-fatal MI of 0.71 (95% CI, 0.58 - 0.87), non-fatal stroke of 1.93 (95% CI, 1.01 - 3.68), and 30 day mortality of 1.33 (95% CI, 1.03 - 1.73).[1] This is not surprising given the large sample size of POISE relative to the other metoprolol studies included in the analysis. A sensitivity analysis was performed to assess the influence of POISE data on the metoprolol outcomes. With POISE data excluded, the odds ratio for non-fatal MI with metoprolol treatment was 0.91 (95% CI, 0.52 - 1.60) p = 0.747, for non-fatal stroke was 2.23 (95% CI, 0.66 - 7.54) p = 0.196, and for short term mortality was 1.11 (95% CI, 0.61 - 2.03) p = 0.340. For all short term outcomes, metoprolol data outside the POISE trial are completely inconclusive. Similarly, no significant evidence of protection from perioperative MI or increased risk of short term adverse events is seen in the other β-blocker group.

To examine the association of perioperative β-blocker initiation and long term mor-



Results from trials included in the Bouri *et al*. analysis stratified by medication

	30 day mortality	Stroke	MI
All other β-blockers	1.09 (0.45, 2.66)	Not estimable	0.97 (0.06, 14.77)
Metoprolol	1.28 (1.00, 1.63)	1.66 (1.05, 2.92)	0.72 (0.59, 0.88)
Metoprolol data without POISE	1.34 (0.57, 1.85)	1.55 (0.52, 4.61)	0.90 (0.51, 1.59)

Figure 3.2: Results of meta-regression analysis of long term (> 6 month) mortality with Eurocare accounting for drug.

tality risk (> 6 months), a second analysis was performed. Figure 2 presents the odds ratios associated with this analysis, both by medication and all β-blockers combined. No increase in mortality is seen with metoprolol treatment when long term outcomes are measured, however, this conclusion is only based on the results of one trial, the DIPOM study. Interestingly, in the stratified analysis, a statistically significant protective effect, odds ratio of 0.50 (95% CI, 0.26 - 0.95) p = 0.034, is seen among β-blockers other than metoprolol. This suggests perioperative β-blockade with agents other than metoprolol is of long term benefit to the patient. Which β-blockers provide this benefit to patients remains unclear.

A likelihood ratio test was performed to test the hypothesis that the relative risk of long term mortality was the same in the metoprolol group versus the other β-blocker group, resulting in $\chi^2(1)$ = 3.875 (p = 0.049). Since this test assesses possible heterogeneity due to the effect of medication, it is recommended by the Cochrane Handbook for Systematic Reviews of Interventions that this p-value be compared to a cutoff of 0.10 instead of the traditional 0.05.[21] This indicates evidence of a statistically and clinically significant difference in the effect of metoprolol versus that of other β-blockers on long term mortality.

3.3.5 Discussion

Previously published observational data have suggested that the risk profile associated with perioperative initiation of metoprolol is significantly different from that of other β-

blockers.[3-5] The likelihood ratio test for the equality of the odds ratio of long term mortality for metoprolol versus other β-blockers presented here provides randomized controlled trial evidence of these differences. The statistical significance of this test combined with the clinically meaningful difference in the estimated effects of the respective medications (OR = 1.03 for metoprolol and OR = 0.50 for other β-blockers for long term mortality) suggests that the results of any mortality analysis that relies on data from patients receiving metoprolol are not likely to be generalizable to patients receiving other β-blockers. Virtually all randomized controlled trial evidence for an increased risk of stroke and short term mortality with metoprolol treatment is derived from POISE trial data. Given these results, there is limited statistical justification for generalizing these outcomes to dosing schemes or protocols other than those employed by POISE. Although only one study of metoprolol had the requisite follow-up time for inclusion in our long term mortality meta-analysis, no association between metoprolol and increased mortality is evident in these data. This suggests a possible crossing-hazards phenomenon.

Randomized controlled trial data on β-blockers other than metoprolol are sparse but do not currently support the hypothesis that these other β-blockers are associated with decreased incidence of perioperative MI or with increased perioperative stroke or short term mortality. The results of the long term mortality meta-analysis presented here demonstrate that as a group, other β-blockers may show a long term protective effect with regard to mortality. Similarly, a recent large observational study found that perioperative use of atenolol was associated with decreased risk of one year mortality compared to metoprolol.[5] Additional randomized controlled trials utilizing β-blockers other than metoprolol focused on both perioperative adverse events and a more comprehensive mortality endpoint need to be performed to assess this effect fully. Ideally these trials would include a full survival analysis, providing estimations of Kaplan-Meier survival curves.

A possible challenge in designing future trials in this field is evident in a statement made by Dr. P.J. Devereaux suggesting that groups at high risk for perioperative adverse cardiac events may receive less benefit from perioperative β-blockade than low-risk groups. Per Dr. Devereaux, this trend was seen in the POISE data but did not reach statistical significance.[23] If this suggested trend represents a difference in the effect of β-blockers among different risk groups, this might imply that previously utilized statistical models, including those presented in this work, are over-simplified. It is possible that perioperative β-blocker initiation provides benefits to one population risk stratum and harm to others. More complicated statistical models accounting for effect modification by baseline risk would need to be employed, and much larger sample sizes than those previously outlined would likely be needed. The current study is limited by the small number of published trials that sufficiently characterize long term outcomes, making it difficult to adequately model time dependent effects of treatment on mortality. It is also limited by the paucity of randomized controlled data pertaining to β-blockers other than metoprolol. Due to this limitation, it was reasonable to analyze these β-blockers as one group. However, this prevents us from drawing conclusions about any individual medications risk profile or mortality effect. Additionally, in our long term mortality analysis we included the EUROCARE[14] study, which utilized carvedilol. It is possible that the effect of carvedilol may be somewhat different from that of selective β-blockers with respect to perioperative risk modification secondary to its mild 1 effects. The authors are unaware of any evidence that this is the case. It should be noted however, that no deaths were observed in the treatment arm of the EUROCARE study. Therefore, statistically this studys influence on the long term mortality analysis was primarily by increasing the precision of the control/placebo groups risk estimate. The controversy surrounding initiation of perioperative β-blockers is far from settled. The POISE trial is currently the primary source of randomized controlled trial data on this subject. It provides the only significant evidence that treatment may be harmful; however, support for the notion that these results may not be generalizable to other β-blockers and dosing schemes is mounting. Recently made claims that 10,000 iatrogenic deaths per year would be prevented by abstaining from initiation of perioperative β-blockade are premature.[9] Results of the analysis presented here demonstrate that perioperative initiation of β-blockers other than metoprolol may actually save patient lives when a more comprehensive mortality endpoint is employed, in addition to any potentially favorable effects on non-fatal perioperative MI rates. Additional randomized controlled trial data are greatly needed to adequately address treatment efficacy and safety of perioperative initiation of β-blockers other than metoprolol.

References

1. Devereaux, P.J., et al., Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. Lancet, 2008. 371(9627): p. 1839-47.

2. Dunkelgrun, M., et al., Bisoprolol and fluvastatin for the reduction of perioperative cardiac mortality and myocardial infarction in intermediate-risk patients undergoing noncardiovascular surgery: a randomized controlled trial (DECREASE-IV). Ann Surg, 2009. 249(6): p. 921-6.

3. Ashes, C., et al., Selective beta1-Antagonism with Bisoprolol Is Associated with Fewer Postoperative Strokes than Atenolol or Metoprolol: A Single-center Cohort Study of 44,092 Consecutive Patients. Anesthesiology, 2013. 119(4): p. 777-87.

4. Mashour, G.A., et al., Perioperative Metoprolol and Risk of Stroke after Noncardiac Surgery. Anesthesiology, 2013.

5. Wallace, A.W., S. Au, and B.A. Cason, Perioperative beta-blockade: atenolol is associated with reduced mortality when compared to metoprolol. Anesthesiology, 2011. 114(4): p. 824-36.

6. McFalls, E.O., et al., Predictors and outcomes of a perioperative myocardial infarction following elective vascular surgery in patients with documented coronary artery disease: results of the CARP trial. Eur Heart J, 2008. 29(3): p. 394-401.

7. Mangano, D.T., et al., Effect of atenolol on mortality and cardiovascular morbidity after noncardiac surgery. Multicenter Study of Perioperative Ischemia Research Group. N

Engl J Med, 1996. 335(23): p. 1713-20.

8. Wallace, A., et al., Prophylactic atenolol reduces postoperative myocardial ischemia. McSPI Research Group. Anesthesiology, 1998. 88(1): p. 7-17.

9. Bouri, S., et al., Meta-analysis of secure randomised controlled trials of beta-blockade to prevent perioperative death in non-cardiac surgery. Heart, 2013.

10. Bolsin, S., M. Colson, and A. Marsiglio, Perioperative beta blockade. Bmj, 2013. 347: p. f5640.

11. Blum, K., Misconduct May Take Bloom Off -Blocker Rx, in Anesthesiology News. 2013, McMahon Publishing: New York, NY.

12. O'Riordan, M., Perioperative Beta-Blocker Controversy Begins Again With New Meta-analysis, in Heartwire. 2013, Medscape.

13. Bayliff, C.D., et al., Propranolol for the prevention of postoperative arrhythmias in general thoracic surgery. Ann Thorac Surg, 1999. 67(1): p. 182-6.

14. Serruys, P.W., et al., Carvedilol for prevention of restenosis after directional coronary atherectomy : final results of the European carvedilol atherectomy restenosis (EUROCARE) trial. Circulation, 2000. 101(13): p. 1512-8.

15. Brady, A.R., et al., Perioperative beta-blockade (POBBLE) for patients undergoing infrarenal vascular surgery: results of a randomized double-blind controlled trial. J Vasc Surg, 2005. 41(4): p. 602-9.

16. Juul, A.B., et al., Effect of perioperative beta blockade in patients with diabetes undergoing major non-cardiac surgery: randomised placebo controlled, blinded multicentre trial. Bmj, 2006. 332(7556): p. 1482.

17. Yang, H., et al., The effects of perioperative beta-blockade: results of the Metoprolol after Vascular Surgery (MaVS) study, a randomized controlled trial. Am Heart J, 2006. 152(5): p. 983-90.

18. Neary, W.D., et al., Lessons learned from a randomised controlled study of perioperative beta blockade in high risk patients undergoing emergency surgery. Surgeon, 2006.

4(3): p. 139-43.

19. Zaugg, M., et al., Adrenergic receptor genotype but not perioperative bisoprolol therapy may determine cardiovascular outcome in at-risk patients undergoing surgery with spinal block: the Swiss Beta Blocker in Spinal Anesthesia (BBSA) study: a double-blinded, placebo-controlled, multicenter trial with 1-year follow-up. Anesthesiology, 2007. 107(1): p. 33-44.

20. Kenneth J. Rothman, S.G., Timothy L. Lash, Modern Epidemiology. Third ed. 2008, Philedelphia, PA: Lippincott Williams and Wilkins.

21. Julian Higgins, S.G., ed. Cochrane Handbook for Systematic Reviews of Interventions. 2008, John Wiley and Sons: West Sussex, England.

22. Rucker, G., et al., Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol, 2008. 8: p. 79.

23. Poldermans, D. and P.J. Devereaux, The experts debate: perioperative beta-blockade for noncardiac surgery–proven safe or not? Cleve Clin J Med, 2009. 76 Suppl 4: p. S84-92.

3.4 Trial Designs for Fuzzy Interventions

This manuscript examines aspects of clinical trial design more generally. In particular it focuses on the fact that large single protocol trials of potentially heterogeneous treatment effects are often justified by power calculations done conditional on an assumed alternative hypothesis. By ignoring the probability of a selected treatment actually achieving that alternative based on the assumed distribution of potential treatment effects, the single protocol design vastly overestestimates its frequency characteristics. By choosing a multiprotocol design the probability of selecting a protocol that meets or exceeds said alternative is improved resulting in more reliable frequency properties.

3.4.1 Likelihood Based Randomization

This work involved the creation of a shiny app which allows investigators to experiment with multiprotocol designs. The most obvious drawback to a fixed randomization, multiprotocol design is that the sample size is split among the various protocols contributing to a lack of precision in the estimated effect of the eventual winning protocol. The simulation allows for two methods to combat this effect both of which rely on an empirical estimate of the likelihood.

In general we know that given data that are distributed f_X, we know that the distribution of the minimum of a sample of size k is given by:

$$f_{X(1)}(x) = kf_X(x)[1 - F_X(x)]^{n-1}.$$

In the simulation built into the shiny app for this project, the $log(RR)$ associated with potential treatments is assumed to be normal and its mean and variance are estimated from the current estimates, $log(\hat{RR})$, from each arm of the study. Given this estimate $\hat{F}_X$ we can calculate a new estimate of $\hat{f}_{X(1)}$ after each study participant. After an initial run-in period during which the randomization is fixed, have a low likelihood of being $x_{(1)}$ are eliminated. Here “low” is a user-defined likelihood ratio compared to the most likely arm. After the run-in period the simulation adopts adaptive randomization in each experimental arm with the randomization probability to a given arm being decided by its likelihood ratio. It has been demonstrated using the app that this method is much more efficient than fixed randomization. The app allows the user to specify the size of the run-in period without constraint. Thus the user may choose to set the run-in size to n corresponding to fixed randomization or to 0 corresponding to fully adaptive randomization. Similarly by setting the threshold likeliood value for elimination after the run-in period to be large, the user can prevent any arms from being discarded.

3.5 Manuscript: Clinical Trial Design for Fuzzy Interventions

3.5.1 Introduction

In clinical trials it is common practice to attempt to reduce extraneous sources of variation in order to achieve the scientific ideal, the realization of a treatment and control group that are balanced except for the administration of a well-defined intervention. Designing studies according to this principle gives increased credibility to a trials conclusions, and provides the foundation for making claims that an intervention results in causal benefit. When several trials are performed to evaluate a common intervention in a fixed population it is reasonable to combine their results via meta-analysis to arrive at an estimate of the common effect associated with that intervention, which is a weighted average of the effects observed across the studies. However, it is often the case that there is an idea for an intervention that could provide clinical benefit for patients, but there are many possibilities for how the intervention could be implemented. For example, interventions involving the administration of medication require the investigator to apriori select a particular medication from a drug class, identify the population likely to derive benefit from the intervention, decide on an optimal dosage and specify a time frame over which the drug is to be administered. All of these study parameters impact the efficacy and safety of the intervention potentially resulting in a distribution of treatment effects that could possibly be observed based on how these study parameters are specified. A situation where there is uncertainty about the optimal intervention is henceforth referred to as a fuzzy intervention. The variability induced by the selection of these parameters is most often ignored outside of investigations of novel drugs.

Throughout this manuscript the example of the safety of perioperative beta blockade for the prevention of myocardial ischemia (MI) will be used as an example. This example was chosen because the issue of this treatments safety is still undecided despite multiple clinical trials involving a large number of participants. The idea behind this intervention is that administering a beta-blocker prior to a surgical procedure might reduce the incidence of MI by causing a reduction in the patients heart rate, and consequently reducing the amount of oxygen required in the heart tissue. A well-defined scientific inquiry regarding safety that one could study might be: Does the administration of 100mg of extended-release metoprolol administered beginning one day prior to surgery and continuing for up to 8 days after surgery increase mortality in patients over 39 years of age, with diabetes, who are undergoing major non-cardiac surgery?[1] However, it is not at all clear that data collected from a study using this protocol speaks directly to the fuzzy intervention question: Can beta-blockers be safely used in the perioperative period in a way that reduces the incidence of MI?

When the range of potential effects from a fuzzy intervention extends from clinically meaningful benefit to clinical deficit based on the choice of protocol, the usual synthesis of data providing a single estimate of a pooled treatment effect is unsatisfying. The average treatment effect is not of any particular interest when it is not the optimal treatment. The scientific question of interest often revolves around the most efficacious treatment that does not compromise safety, which could be far from the average treatment effect.

The marked difference in the clinical questions being asked in the well-defined intervention versus the fuzzy intervention underscores the need to approach these problems in different ways. This manuscript seeks to expose the danger of analyzing trials involving a fuzzy protocol using traditional techniques. A series of simulation studies then assess how embracing a variety of protocols in the administration of fuzzy intervention trials leads to more precise effect estimates and allows the investigator to more thoroughly address the true object of their inquiry, characterization of the optimal treatment protocols.

3.5.2 Motivating Example

The safety of administering a beta-blocker before surgery in an attempt to reduce cardiac oxygen demand in order to prevent cardiac ischemia has been an ongoing controversy in the medical literature for many years. The intervention involves the administration of medication, often orally, before and after non-cardiac surgery. In each trial investigating the safety and efficacy of this intervention the investigators chose a particular beta-blocker to study, a surgical population, a dosing scheme, and a timeframe over which the drug was to be administered.

At the heart of the controversy is the POISE trial.[2] The POISE trial is by far the largest study of perioperative beta-blockade to date (n=8351). The POISE trial utilized extended-release metoprolol. The chosen dosing scheme began with 100mg 2-4 hours prior to surgery followed by up to 300mg in the first 18 hours postoperatively. Study medication was continued for 30 days postoperatively. The trial concluded that while the treatment appeared to be efficacious for the prevention of myocardial infarction it was also associated with an increase in risk of stroke and mortality.

Many clinicians were unconvinced by the POISE results due to the large dosage employed.[3] In addition, several large observational studies have since suggested that metoprolol has a significantly worse safety profile when compared to other beta-blockers administered in a similar fashion perioperatively.[4-6] Many meta-analyses were published to try to ascertain the strength of evidence for the safety of the intervention, but due to POISEs disproportionate size any analysis that included POISE would essentially be reporting its result. The analysis by Bouri et.al.[7] is an example. In their analysis 77.8% of the weight was placed on POISE with 92% of the weight coming from studies involving metoprolol. It is easy to see how the resulting estimate could be interpreted as largely representing the average effect of high-dosage metoprolol treatment rather than deciding the safety of the optimal protocol. Here despite the large number of patients exposed to the treatment over the included studies (n=5264 exposed to some perioperative beta blocker) the true safety of the optimal protocol for beta-blockade remains undecided[8] with the only potentially undisputed evidence being against the use of metoprolol.

3.5.3 Proposed Method

In situations where an investigator is faced with a fuzzy intervention it can be advantageous to examine a number of different protocols as opposed to conducting one large trial according to a single protocol. Fuzzy scientific questions dictate that it is necessary to estimate the distribution of possible treatment effects rather than just the average of the distribution. Each protocol considered adds an estimate of a single data point drawn from the distribution of possible treatment effects measured with some error that depends on the number of patients assigned to that protocol. Increasing the number of patients studied under a given protocol improves the precision with which the single data point is estimated. The estimates from these protocols can then be used for estimation of the distribution of possible effects. When trying to decide if a treatment provides a substantial clinical benefit there is a tradeoff to be made between the precision with which you measure a given protocols efficacy and including additional protocols, which may have superior efficacy or safety.

The remainder of this manuscript details how investigators that intend to conduct large studies should choose the number of protocols to examine. By optimizing the number of protocols under examination relative to the number of subjects exposed to each protocol it will be shown that the probability of identifying a more optimal treatment protocol will increase.

3.5.4 Fuzzy Interventions

When conducting a clinical trial of a fuzzy intervention it is necessary to reexamine the ordinary metrics associated with clinical trial design. For example, what is the null hypothesis when a treatment effect is heterogeneous and therefore unlikely to truly have absolute zero effect? Because type I errors are defined by falsely claiming efficacy when the null hypothesis is true, being in a situation where the null is unlikely to ever be true necessitates a redefinition. Defining properties that are desirable for a trial to have will enable the investigator to optimize the number of protocols to consider in a given study.

Type I Errors: Type I errors are poorly defined when the treatment under consideration is heterogeneous. However, it is still desirable to make certain that a multiple-protocol trial investigating a fuzzy intervention will not declare a treatment efficacious if the effect is either of insignificant magnitude or worse, deleterious. This deviation from the traditional setting where type I errors are well-defined requires a different approach to what is considered a null effect.

Perhaps the most straightforward quantity related to type 1 error rate in a multi-arm trial of a fuzzy intervention is the probability that a protocol deemed optimal by the trial actually has a deleterious effect. A related quantity that is perhaps less objective is the probability that a protocol, which is deemed optimal by the trial actually has a clinically insignificant benefit or worse. As the latter of these quantities is both situation specific and subject to personal opinion, for the duration of the manuscript we will restrict discussions to the former. This idea has been previously introduced in social science where it is described as type S error [9].

Unlike traditional type I errors that are conditional on the null hypothesis of no effect, the type S error rate depends on the true distribution of heterogeneous treatment effects. In order for the error to be made, a deleterious effect must be chosen during the initial protocol selection. The arm of the study corresponding to the deleterious effect must then outperform all the other arms. Finally, that same arms estimate of treatment effect must have sufficient precision such that its corresponding uncertainty interval does not contain null or deleterious values after an appropriate analysis has been conducted. This error rate can be estimated via simulation under a variety of foreseeable distributions of potential treatment effects. When the study is completed the investigator can then estimate a distribution of potential treatment effects to be used in simulation to estimate the post-hoc type S error rate of the study in a way that naturally accounts for multiple comparisons.

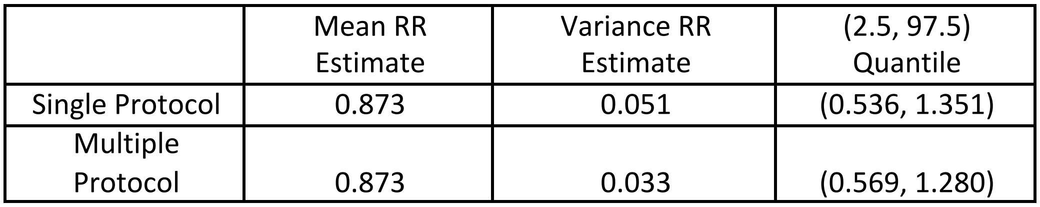
Power: Like type I errors, power also experiences a minor change in the fuzzy intervention setting. In the classical framework the power of a trial is the probability that the study is able to reject the null hypothesis given that the alternative is true. When dealing with heterogeneous treatment effects, however, it can become either unlikely or impossible that a treatment has an absolutely null effect depending on the distribution of potential effects. For example, one might think that giving someone an analgesic for a headache would have no effect on 30-day mortality. However, there is always a very small chance the patient will have an anaphylactic reaction. In reality there are few, if any, interventions with absolutely no risk. It does not therefore seem inappropriate to assume that every selected protocol will have some non-null, if potentially infinitesimal, effect on the outcome of interest.

As a result of the assumption of non-nullity among the protocols, the usual definition of power reduces to simply the probability that the investigator is able to estimate the optimal protocol with sufficient precision such that its uncertainty interval limit excludes the null. This definition can be further restricted for the purposes of trial design to exclude cases when the optimal protocol is estimated to be detrimental, a quantity we will refer to as fuzzy power.

Again with power we see that the fuzzy power is dependent on the true distribution of treatment effects. In order to be declared likely to be beneficial at a given level of precision, the estimate of the protocol deemed optimals effect must be sufficiently small. The probability with which this happens depends on the probability of selecting protocols with large benefits as well as the number of independent protocols selected for examination. This probability can also be simulated under a variety of potential treatment effect distributions during study design to ensure a cost-efficient design is selected.

3.5.5 Simulations

Several simulations were conducted to illustrate the benefits of large clinical trials using multiple protocols. These simulations are similar to those that would be done during the



	Mean RR Estimate	Variance RR Estimate	(2.5, 97.5) Quantile
Single Protocol	0.873	0.051	(0.536, 1.351)
Multiple Protocol	0.873	0.033	(0.569, 1.280)

Table 3.4: Results from mean estimation from a single treatment protocol randomized trial vs. a multiple protocol approach demonstrating increased precision of mean RR using the multiple protocol approach.

trial design phase. For practical purposes each simulation must assume a distribution of potential treatment effects for the intervention being considered. The SHINY app accompanying this paper can be used to conduct additional simulations, and allows the user to specify many aspects of trial design. The following simulations can be reproduced using the app under a variety of different conditions. Simulation studies were conducted using R version 3.2.0. The accompanying web tool was developed using Shiny 0.12.

Example 1: Evaluating Estimates of the Mean Treatment Effect In the classical approach to large trials it is assumed that there is either a single effect of treatment, or if treatments are believed heterogeneous, that interest lies with the average effect. The purpose of this example is to compare the accuracy and efficiency of estimating the mean treatment effect when using a large study with one treatment protocol versus a study of identical size that utilizes multiple protocols. In the case of the multiple-protocol design, a random study size between 40 and 200 was generated for each of the five random protocols. The size of the large, single-protocol study was equal to the size of the five small studies combined.

This simulation takes a given average risk ratio for the treatment effect, a standard error of the observed baseline rate that might be observed among sites participating in the study, and a standard error representing the degree of heterogeneity in treatment effect. Baseline Event Rate $N(\mu = 0.15, \sigma = 0.05)$

Protocol Specific Risk Ratio $\sim logN(log(\mu) = log(0.85), log(\sigma) = 0.025)$

Having specified these study parameters, data was generated for the trials and a risk ratio is calculated directly for the large study and by way of a DerSimonian-Laird random effects model for the multi-protocol study. The simulation was completed 10,000 times and the operating characteristics of the two approaches were compared. The results of the simulation are given in Table 1. The simulation shows that in addition to providing an estimate of the variance in treatment effect among various possible treatment protocols, the mean RR estimate derived from the multiple protocol approach is substantially more precise than the single protocol approach with similar accuracy. This demonstrates that the multiprotocol approach can be used to estimate the mean of the distribution of potential treatment effects in an efficient manner. This simulation was included for those who may be reluctant to abandon the traditional inference on the mean approach. The main result of the simulation is that the additional information provided from the multiple protocol approach will often produce a superior result even using the traditional techniques of inference.

The difference in the efficiency of the estimate varied with the magnitude of heterogeneity of the treatment effect. This experiment was completed at a variety of different potential dispersions ranging from 0 to 0.4. The resulting differences in the variance are reported in figure 1.

Example 2: The next simulation will mimic the proposed design of a large clinical study with potentially heterogeneous treatment effects between a myriad of protocols that are clinically reasonable. For reasons of convenience, this simulation will be modeled after POISE and explore how a multiple protocol approach would have potentially provided a more robust body of evidence on the safety of perioperative beta blockade with respect to patient mortality.

The first step in designing our hypothetical trial will be to decide on what feasible distributions of potential outcomes might look like. Several distributions could be chosen and tested, however, often times there have been many other smaller trials or observational studies that can be utilized to give the investigators a good initial guess at the distribution.

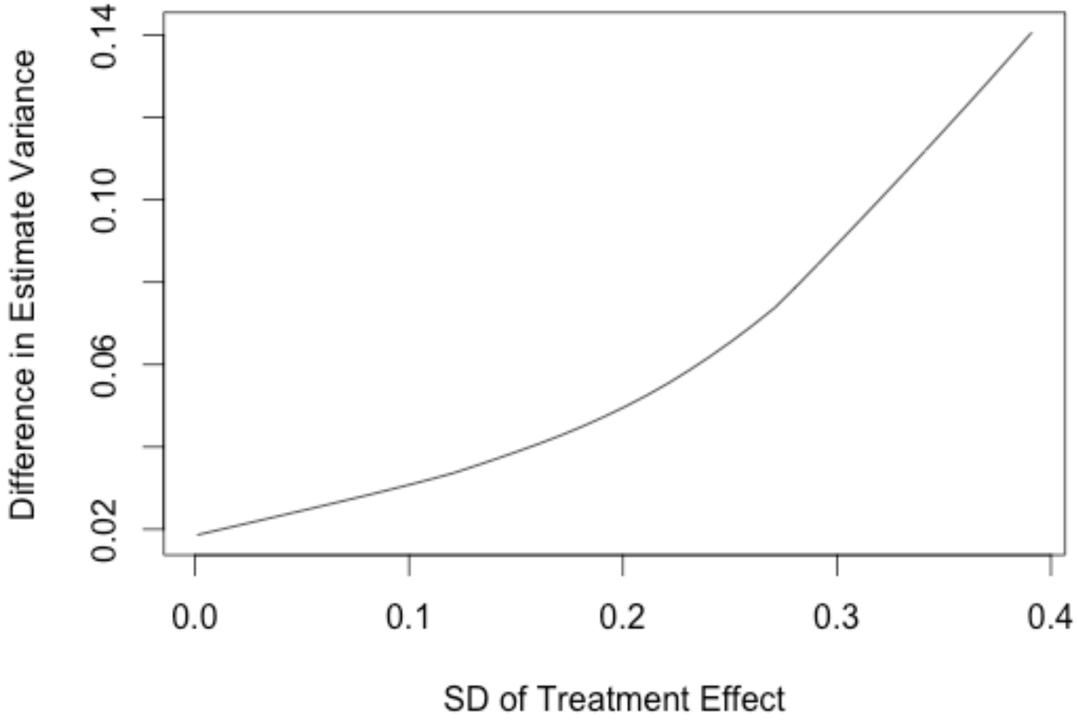


Figure 3.3: Difference in variance of the mean treatment effect for single protocol vs multiprotocol design as a function of the SD of the log risk ratio of potential treatments.

The investigator could then conduct simulation studies to assess the operating characteristics of the trial with a variety of numbers of protocols being tested. For this simulation a normal distribution with a mean of zero and standard deviation of 0.05 was chosen as the distribution of potential relative risks corresponding to the effects of treatment. This corresponds to the treatment having a RR between 0.9 and 1.1 depending on the particular chosen protocol.

Since one of the goals of the trial is to identify beneficial protocols, the first metric we consider is the mean relative risk of the protocol that the study identifies as optimal. In other words, given a distribution of treatment effects it is desirable to select the design, which on average, results in a protocol with the most beneficial relative risk. After conducting many simulations with 1, 2, 3, etc. protocols, one can simply construct a scree plot to decide how many protocols to investigate at a given sample size and design. The sample size and design have to be considered simultaneously because they will both influence the studys fuzzy power.

In order to enhance the probability of identifying and rejecting a beneficial treatment when one exists, it may be desirable to use an adaptive randomization scheme. For the purposes of this example, once the number of protocols for the trial has been selected an initial run-in trial is performed in which each experimental arm of the study has an equal chance of being randomized to. After the initial run in period, arms that had a low relative probability of being best were discarded. The study is completed using the remaining arms for which the probability of being randomized to a given arm is proportional to the relative likelihood of that arm being best, which is updated after each patient finishes the study. Throughout the study, patients are twice as likely to be randomized to control as they are to be randomized to a given experimental arm. In this simulation we used a run-in of 2000 patients, and a relative probability of 1/10 for arms being discontinued after the run-in. This design selection process can be iterated over at a variety of different sample sizes to determine the sample size required in order for the type S error rate and fuzzy power

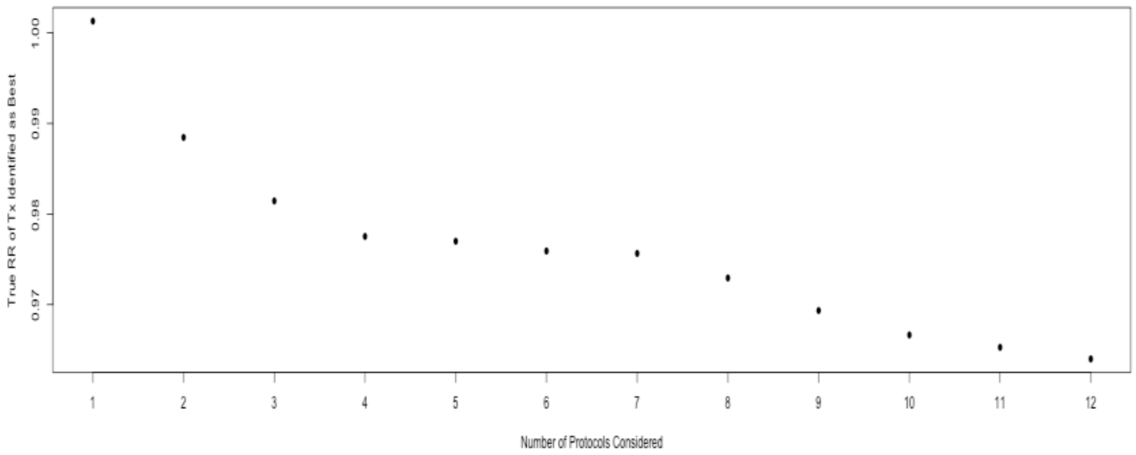


Figure 3.4: Mean true RR of the protocol identified as best as a function of the number of protocols being examined.

to meet the investigators needs. For the purpose of this simulation, the sample size was considered fixed at the sample size used in the POISE trial, 8351.

The scree plot and characteristics of the resulting study termed best for this simulation are given in Figure 2 and 3. As evidenced by the scree plot, under the distributional assumption for potential treatment effects on average better treatments are found as more protocols are considered considered. The error rate plot demonstrates that increasing the number of protocols concurrently drives the probability of the best protocol being detrimental down. Correspondingly the fuzzy power of the study is increasing as selecting more sets of parameters to investigate results in a higher probability of a more beneficial one. In fact, if a beneficial protocol exists, it is nearly 10 times as probable to be discovered in study with twelve experimental arms as in a study with just a single experimental arm.

3.5.6 Discussion

The primary purpose of this manuscript is to emphasize the importance of designing a trial in order to answer a specific scientific question. In medicine, the scientific questions we face are often multifaceted, i.e. What treatment is best, and how well can I expect it to perform? However, altogether too often when faced with a variety of potential interventions

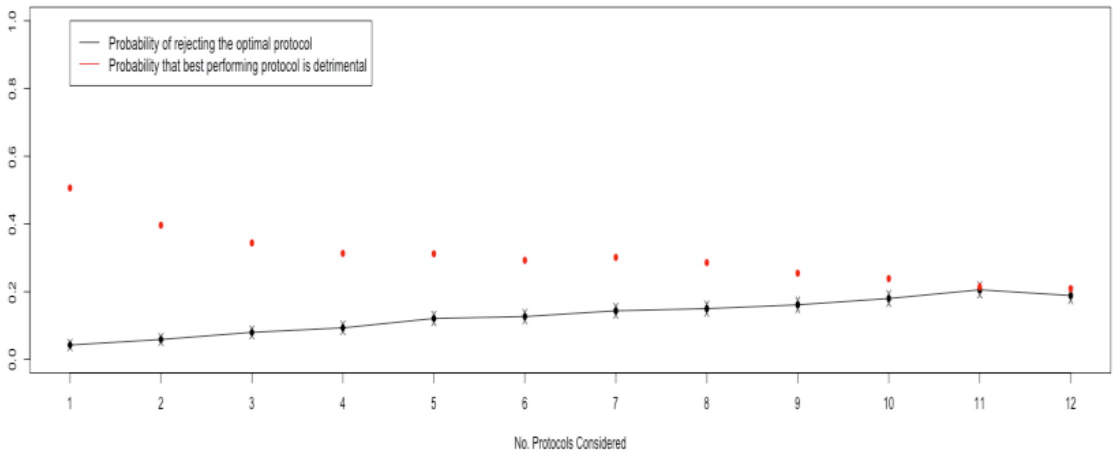


Figure 3.5: Red Probability that the best performing protocol is detrimental regardless of whether it was rejected or not. Black- Probability the best performing protocol is deemed beneficial.

investigators often lose sight of the first question in favor of the second.

The current mindset in clinical trial design is to attempt to reduce variation in the administration of an intervention as much as is possible. However, this goal while laudable in some cases where the protocol is well agreed upon is rarely achieved in practice as most trials can be critiqued on some point where bias could enter the estimation process. By relaxing the restrictions on the trial and embracing multiple protocols, embracing structured variation, and adopting appropriate multi-level modeling [10] or empirical Bayes estimation a trial has the opportunity to improve its operating characteristics while also addressing the scientific question that is most clearly relevant, Can this intervention benefit patients and if so how should it be implemented? Whether and to what degree this improvement is achievable depends on the specifics of potential interventions and the size of the trial.

The results of the first simulation demonstrate that the classical approach to evaluating complex, multifactorial clinical interventions, namely the large single protocol clinical trial, can be substantially improved upon by designing multi-protocol studies that collect more diverse information. It shows that even when an analyst restricts themselves to a traditional inference on the mean approach, the mean estimate is often improved by examining multiple protocols.

The second simulation and the accompanying app demonstrate how a multiple protocol clinical trial might be designed, and is meant to highlight the advantages of this approach. In general, the probability of identifying a deleterious protocol is non-increasing as the number of arms under investigation increases and the fuzzy power is non-decreasing. Although these properties hold generally, depending on the assumed distribution of treatment effects, substantial benefits may be seen from increasing the number of protocols.

The last consideration with regard to designing clinical trials of fuzzy interventions is an ethical one. In the classical approach of a single protocol design, millions of dollars are often spent and thousands of patients put at risk to estimate the effect of what is essentially a random draw from the distribution of potential treatment effects. Although one might argue that the subject matter experts conducting the trial are more likely than not to select a good protocol, one does not have to look far in the medical literature to identify a proposed treatment in which many trials looked at a myriad of protocols, each of which was selected by an expert in the field. Indeed as we have mentioned the evidence that the medication selected for use in the POISE trial, metoprolol, is significantly less safe than other beta blockers when employed for perioperative beta-blockade is mounting. Despite having some of the worlds foremost experts on the intervention deciding on the protocol, they still chose one that seems to have the highest associated stroke and mortality rates. The risk to human life as well as the expense obligates investigators to design a trial that is most probable to identify a protocol from which society will derive some benefit, if one exists. In many cases, that ethical obligation mandates a multiple protocol approach.

When the uncertainty associated with the selection of a trial protocol from a family of potential protocols is included in the simulation the advantage to multiprotocol approaches becomes apparent. Single protocol designs are often justified by power calculation in which an alternative assumption is made that may only have a small probability of being true. Improving this probability of achieving the alternative assumption results in improved frequency characteristics.

References

1. Juul, A.B., et al., Effect of perioperative beta blockade in patients with diabetes undergoing major non-cardiac surgery: randomised placebo controlled, blinded multicentre trial. Bmj, 2006. 332(7556): p. 1482.

2. Devereaux, P.J., et al., Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. Lancet, 2008. 371(9627): p. 1839-47.

3. Poldermans, D. and P.J. Devereaux, The experts debate: perioperative beta-blockade for noncardiac surgery–proven safe or not? Cleve Clin J Med, 2009. 76 Suppl 4: p. S84-92.

4. Ashes, C., et al., Selective beta1-Antagonism with Bisoprolol Is Associated with Fewer Postoperative Strokes than Atenolol or Metoprolol: A Single-center Cohort Study of 44,092 Consecutive Patients. Anesthesiology, 2013. 119(4): p. 777-87.

5. Mashour, G.A., et al., Perioperative Metoprolol and Risk of Stroke after Noncardiac Surgery. Anesthesiology, 2013. 119: p. 1340-6.

6. Wallace, A.W., S. Au, and B.A. Cason, Perioperative beta-blockade: atenolol is associated with reduced mortality when compared to metoprolol. Anesthesiology, 2011. 114(4): p. 824-36.

7. Bouri, S., et al., Meta-analysis of secure randomised controlled trials of beta-blockade to prevent perioperative death in non-cardiac surgery. Heart, 2013. 100(6): p. 456-64.

8. Wijeysundera, D.N., et al., Perioperative beta blockade in noncardiac surgery: a systematic review for the 2014 ACC/AHA guideline on perioperative cardiovascular evaluation and management of patients undergoing noncardiac surgery: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines. J Am Coll Cardiol, 2014. 64(22): p. 2406-25.

9. Gelman, A. and J. Carlin, Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspect Psychol Sci, 2014. 9(6): p. 641-51.

10. Gelman, A.H.J.Y.M., Why we (Usually) Don’t Have to Worry About Multiple Comparisons. Journal of Research on Educational Effectiveness, 2012. 5: p. 189-211.

Chapter 4

Probabilistically Calibrated Support Intervals

4.1 Introduction to the R Package

Previous to the release of *supportInt*, there was no package in R focused on the calculation of likelihood support intervals. *supportInt* works for a wide range of common data types including binomial, poisson, normal, linear models and generalized linear models. Using only its basic functionality, it provides a much needed tool for likelihood based inference.

The package also supports additional functionality that allows bootstrap estimation of the frequency properties of support intervals for the three basic data types. It uses a novel, sophisticated bootstrap scheme described in the following manuscript to overcome instabilities in the coverage probability for binomial confidence intervals of small *n* samples. It also provides the *calibSI()* function, which will estimate the minimum support level required to acheive the desired frequency properties.

The support intervals generated by *supportInt* address a major concern that many philosophers and statisticians have with likelihood based inference, that it is not probabilistic. Many consider probability based inference a stronger form than purely likelihood based inference, although the strongest argument for this position appears to be an obscure quote from R. A. Fisher giving his personal opinion on the matter. Even still, this position has been a major obstacle to the use of likelihood methods. Through bootstrap probabilistic calibration, these likelihood methods achieve a probabilistic interpretation which invalidate this common objection.

4.2 Manuscript: Probabilistically Calibrated Support Intervals with *supportInt*

4.2.1 Abstract

The supportInt package calculates likelihood support intervals in a variety of model contexts. The package calculates support intervals for outcomes from binomial, poisson, and normal distributions as well as for regression coefficients from lm() and glm() models. The binomial, poisson, and normal functions also support the use of a novel bootstrap technique to estimate frequentist coverage rates for these support intervals allowing a probabilistic interpretation.

4.2.2 Background

One of the undertakings in statistics is the estimation of sets of plausible values for model parameters. Regardless of what these parameters represent, much of statistical practice revolves around the derivation of sets of potential values based on the evidence contained in the data. In the case of likelihood-based statistical inference, the preferred sets are referred to as support sets, or when composed of contiguous values, support intervals.

Other types of uncertainty intervals (confidence and credible intervals) define a set of potential parameter values that will contain the true parameter with some frequency or probability without specifying any particular requirement that a potential parameter value must satisfy for inclusion in the interval. Because these intervals are defined by a property of the set of values rather than a property of the individual values within the set, it is not uncommon for confidence and credible intervals to include values with relatively poor evidential support. Unlike other probabilistically-based intervals, support intervals are composed of potential parameter values that achieve some specified level of evidential support from the data, however the resulting set lacks any probabilistic interpretation. This deficiency has lead many analysts to prefer confidence type procedures, because they view probabilistic inference as being stonger than pure likelihood inference. The probabilistic calibration of support intervals proposed here and implemented in the *supportInt* package adjusts the level of evidential support so that the resulting support interval has the probabilistic interpretation that has been historically lacking.

At present there are no R packages that calculate likelihood support intervals for basic types of data. Two packages for the analysis of genetic data calculate support intervals *qtl* (17) and *mpMap* (18), but only for a very specific statistic used in genetics. *supportInt* fills this key gap for likelihoodists who use R.

4.2.3 The Likelihood Function

The likelihood function is the primary data summary that underlies a large portion of statistical theory. Given independent and identically distributed observations $x_1, \ldots, x_n$ from a probability distribution $f(X|\theta_0)$ where θ_0 is unknown by the observer, the likelihood function is given by:

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta).$$

The likelihood is a function defined over the potential parameter values. At each parameter value the likelihood is equal to the joint density evaluated at the observed data values. The meaning of the likelihood is more apparent when f happens to be a discrete distribution. In this case, the likelihood at a given parameter value, θ, is simply the probability of observing the data when $\theta_0 = \theta$. The potential parameter value that results in the highest likelihood of having observed $x_1, \ldots, x_n$ is called the *maximum likelihood estimator* (MLE), $\hat{\theta}$. When evaluating the evidential support for a potential parameter value its likelihood is often compared to this maximum value, (19). In this way the most well-supported potential parameter value becomes the benchmark against which all other potential values are judged. Therefore we often deal with the normalized likelihood function:

$$\frac{L(\theta|x_1, \ldots, x_n)}{L(\hat{\theta}|x_1, \ldots, x_n)} = \frac{\prod_{i=1}^{n} f(x_i|\theta)}{\prod_{i=1}^{n} f(x_i|\hat{\theta})}.$$

The normalized likelihood function is the likelihood ratio between a given value in the parameter space, and the maximum likelihood over the parameter space. Its values represent the multiplicative change in the evidential support between the evaluated value and the MLE. For example, a normalized likelihood value of $1/8$ at θ' would mean that seeing the observed data if the $\theta_0 = \hat{\theta}$ is eight times better supported than seeing it if $\theta_0 = \theta'$. Given a model, the normalized likelihood is a completely data-based summary of the degree to which various potential values for a parameter are supported by the data.

4.2.4 Support Intervals

In the likelihood paradigm, inferences are based solely on the data without regard to the sample space or any prior information not contained with the observed data, and the normalized likelihood is the primary tool of inference. A $1/k$ support set is defined to be the set of all potential parameter values in the parameter space of θ, Θ, for which:

$$\left\{\theta \in \Theta : \frac{L(\theta|x_1,\ldots,x_n)}{L(\hat{\theta}|x_1,\ldots,x_n)} > \frac{1}{k}\right\}$$

The support set is the set of all potential values of θ under which the data would not have been more than k times better supported by any other value of θ in Θ. When these values are contiguous, as is usually the case when dealing with well-behaved, unimodal likelihoods, the support set is called a support interval and will be referred to as such for the duration of this discussion. Support intervals can be interpreted as the values best supported by the data at the specified level, k.

4.2.5 Properties of Support Intervals

4.2.6 Invariance

Invariance to transformation is a very important property enjoyed by support intervals. Support intervals inherit their invariance from the likelihood function where it can be shown very generally that

$$\frac{L(\theta|x_1,\ldots,x_n)}{L(\hat{\theta}|x_1,\ldots,x_n)} = c \Rightarrow \frac{L(g(\theta)|x_1,\ldots,x_n)}{L(\widehat{g(\theta)}|x_1,\ldots,x_n)} = c.$$

This means that a support interval can be calculated on any scale and subsequently transformed to any other new scale, and the result will be the same as if the interval had been calculated on the new scale. To fully appreciate the magnitude of the benefit of invariance, compare this result to frequentist confidence intervals.

The cases in which an exact 95% confidence interval is available are limited. For cases where an exact interval is not available the nearly universal technique involves invoking some Normal approximation either by appealing the central limit theorem or by finding a transformation, which would make the sampling distribution approximately normal. However, the convergence provided by the central limit theorem can be slow. Some statistics take many thousands of observations before their sampling distributions are an acceptable approximation to normality, if they converge at all. The transformation approach is equally problematic as it supposes that an invertible transformation exists that would make the sampling distribution approximately normal. Presuming such a transformation exists, the analyst then has the unenviable task of figuring out what the proper transformation is. In contrast, due to the invariance property inherited from the likelihood function, deriving a support interval is as simple as calculating the interval on the most convenient scale and transforming it to the desired scale with no approximation required.

4.2.7 Non-exclusivity

Non-exclusivity in this context means that support intervals are by definition related to the level of support a potential value for θ receives from the observed data. Consequently, it is impossible for a support interval to include one value in the support interval while excluding another value with the same or greater evidential support, i.e. normalized likelihood value. The support interval's relationship with the level of evidence observed ensures that any value with the given support is included.

To see that other intervals may fail this criterion, consider a confidence or credible interval constructed for a discrete valued parameter. To maximize the evidential support, the interval is constructed by adding the parameter value with the highest likelihood and evaluating the frequency or posterior probability as appropriate. Values are then added in order of decreasing likelihood until the nominal $1-\alpha$ level coverage is achieved. Now presume this process has been carried out and we currently have an $1-\alpha-\varepsilon$ level interval, but the next two potential parameter values in the sequence have exactly the same likelihood. Adding either of them will give the interval $1-\alpha$ level coverage, but adding both will result in over-coverage. Confidence and credible intervals are defined by a property of the interval rather than individual values. In both cases one value is added to the interval and the other is not. This arbitrary exclusion is forbidden by support intervals.

4.2.8 Nuisance Parameters

Nuisance parameters are parameters of the specified model other than the parameter currently under examination. Likelihood methods are easiest to interpret when the likelihood function depends on a single parameter. This allows the likelihood function to be visualized and it is straightforward to calculate a support interval for a single parameter. When more than one parameter exists, the two most common ways to deal with nuisance parameters are to estimate them or to profile them out. Estimating nuisance parameters results in a likelihood function called an *estimated likelihood*. Although this method works well in some situations it can misrepresent the likelihood when there is a strong joint relationship between the parameter of interest and the nuisance parameters. Profiling out a nuisance parameter simply means that for each value of θ the nuisance parameter is replaced with its MLE conditional on the value of θ. This more conservatively accounts for any relationships between the nuisance parameters and the parameter of interest and tends to behave well in a wide variety of situations.

4.2.9 Support Intervals in the supportInt Package

The supportInt package for R (20) allows calculation of support intervals at the user specified level for binomial, poisson, or normal data directly from their likelihood functions with the *binLikSI(), poisLikSI(),* and *normLikSI()* functions. These functions utilize a root finding algorithm to return support intervals at the user specified level. *supportInt* also utilizes the *ProfileLikelihood* package (21) to provide profile likelihood intervals for coefficients of both lm and glm models with the *lmLikSI()* and *glmLikSI()* functions.

```
library(supportInt)
binLikSI(dat=8, n=10, level=8)
   [1] 0.4877142 0.9667507


poisLikSI(dat=c(4, 4, 3, 5), level=8)
   [1] 2.291555 6.399550


normLikSI(c(4, 3.2, 5.1, 6.8), level=8)
   [1] 2.949658 6.600198



set.seed(10)
```

```
x <- rnorm(50, 0 , 5)
y <- sapply(1:length(x), function(z) 3+.5*x[z]+rnorm(1, 0, 5))
lm.obj <- lm(y~x)

lmLikSI(lm.obj, 8)
       low 1/8   upp 1/8
   x 0.2187735 0.8818569

set.seed(10)
x <- rnorm(50, 0 , 5)
x2 <- rbinom(50, 1, .2)
expit <- function(z) exp(z)/(1+exp(z))
p <- expit(.1+ .4*x+.3*x2)
y <- sapply(1:length(p), function(z) rbinom(1,1, p[z]))
glm.obj <- glm(y~x+x2, family="binomial")

glmLikSI(glm.obj, 8)
         low 1/8  upp 1/8
   x   0.3022526 1.000502
```

4.2.10 Probabilistically Calibrated Support Intervals

Despite having some attractive properties to their credit, support intervals have yet to gain widespread application in statistical practice. Perhaps the single biggest objection to the use of support intervals is that they lack a probabilistic calibration, and whether probabilistic inference is more or less persuasive than likelihood-based inference is hotly debated. The bootstrap makes it possible to have report the evidential support of a probabilistic interval in a wide variety of cases.

Probabilistically calibrating a support interval may require some compromise on the part of both the frequentist and the likelihoodist. To probabilistically calibrate a support interval, the analyst needs to specify the desired frequency properties and then report whatever level of statistical evidence is required to achieve the specified frequency characteristics. In contrast, in traditional likelihood inference, one would choose the level of statistical evidence that was meaningful to them and report the corresponding $1/k$ support interval/s. In the normal case, a $1/6.83$ SI is an exact 95% confidence interval. Royall (19) recommends $1/8$ and $1/32$ for levels that have some interpretability with respect to how they relate to seeing strings of consecutive heads in coin flips, but the numeric values are admittedly arbitrary. Given the somewhat arbitrary nature of the numerical values this may be less of a sacrifice, but since different models and sample sizes will return different levels of likelihood-based support associated with their $(1-\alpha)\%$ CIs, the support levels may not be reported in a consistent way across manuscripts and therefore may be somewhat harder to compare. Similarly, the frequentist will be unable to attain the exact coverage level in some cases due to the definitional constraints imposed by the support interval and is will be forced to settle for a slightly higher or lower coverage.

The aformentioned functions in the *supportInt* package *binLikSI(), poisLikSI(),* and *normLikSI()* will all provide an estimate of the confidence level of the support interval they return if the user specifies the *conf=TRUE* argument. The precision of this estimate can be controlled by increasing or decreasing the number of bootstrap simulations through the *B* argument. In addition, the *calibSI()* function allows the user to specify a desired confidence level for binomial, poisson, or normal observations and will return an approximate confidence interval with the corresponding support level needed to obtain that confidence.

```
binLikSI(dat=8, n=10, level=8, conf=TRUE, B=3000)
   $si
   [1] 0.4877142 0.9667507
   $conf.equiv
   [1] 0.948

poisLikSI(dat=c(4, 4, 3, 5), level=8, conf=TRUE, B=3000)
   $si
   [1] 2.291555 6.399550

   $conf.equiv
   [1] 0.9546667

normLikSI(c(4, 3.2, 5.1, 6.8), level=8, conf=TRUE)
   $si
   [1] 2.949658 6.600198

   $conf.equiv
   [1] 0.9585833

calibSI(dat=8, n=10, family="binomial", conf.level=.95, B=3000)
   $si
   [1] 0.4846169 0.9674700

   $support.level
   [1] 8.318107

   $init.grid
         st.levels cov.st.levels
    [1,]         4      0.8784
```

```
[2,]      6    0.9260
[3,]      8    0.9490
[4,]     10    0.9622
[5,]     12    0.9696
[6,]     14    0.9712
[7,]     16    0.9766
[8,]     18    0.9846
[9,]     20    0.9828
```

4.2.11 Probability Calibrated Gaussian known σ^2

As is often the case when it comes to uncertainty intervals, Gaussian data with known variance is the least troublesome. It can be shown that

$$2\log\left(\frac{L(\theta|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})}\right) \sim \chi^2_1.$$

This result, known as Wilk's statistic, provides the basis for the likelihood ratio test, which can be inverted to give a proper $(1-\alpha)$%CI that is also a SI with

$$\frac{1}{k} = e^{-1/2\chi^2_{1,(1-\alpha)}}$$

level of support, where $\chi^2_{1,(1-\alpha)}$ is the $(1-\alpha)$ quantile of the χ^2_1 distribution. This simple probability calibration holds exactly in this special case, but becomes more general for MLE associated uncertainty intervals under some fairly permissive regularity conditions. As an example, a 95% CI for the Normal mean with known variance is also a $1/e^{-1/2\chi^2_{1,.95}} \approx 1/6.826$ level SI.

4.2.12 Bootstrap Calibrated Non-Gaussian

In the absence of an exact relationship between confidence level and support level enjoyed in the normal case, extra steps will be required to perform probabilistic calibration. It is still possible to calculate inverted likelihood ratio test intervals by either 1) taking a large enough sample to invoke the asymptotic approximation to Wilk's Statistic or 2) by bootstrapping the distribution of the likelihood ratio at the given sample size. However, since statisticians can rarely command a larger sample size based purely on distributional convergence rates, this work focuses on the latter. This technique, performed by Owen (22) using empirical likelihoods is equivalent to inverting a bootstrap corrected likelihood ratio test. The technique involves bootstrapping the distribution of the likelihood ratio by taking bootstrap samples and calculating

$$R^* = \frac{L(\mathbf{x}^*|\hat{\theta}^*)}{L(\mathbf{x}|\hat{\theta})}.$$

$R*$ is the likelihood ratio of the bootstrap data when theta is taken to be the MLE derived from the bootstrap sample vs. when theta is taken to be the MLE of the actual observed sample. Taking the appropriate quantiles of the resulting bootstrap distribution of $R*$ will provide a more robust $1-\alpha$ CI. This procedure performs well in many cases, but in non-Gaussian cases if the sample size is so small that appealing to the Wilk's theorem is impractical it may also be small enough that bootstrapping a non-smooth statistic (quantile) in the tail of a distribution will also perform poorly.

An alternative method employed by *supportInt* is to bootstrap the coverage probability as opposed to bootstrapping the distribution of $R*$ (9). This is similar to Efron's studentized bootstrap t-interval (24) in which the quantile of the t distribution is adjusted until the bootstrap estimate of coverage is nominal. One can similarly use the bootstrap to estimate the coverage probability of a $1/k$ SI and adjust the level, k, until $(1-\alpha)$ coverage is achieved. Since this procedure involves bootstrapping a mean rather than a quantile, it performs better in small samples.

As an example, consider if the data $(5,4,4,2,4)$ are taken to be from a Poisson distribution. Then $\hat{\theta} = \bar{x} = 3.8$. Parametric bootstrap draws are then done from:

$$x^* \sim f(x|\theta = \hat{\theta}) = \frac{e^{\hat{\theta}}\hat{\theta}^x}{x!}.$$

Next, $1/k$ support intervals are calculated from the bootstrap sample $x_1^*, \ldots, x_5^*$, and counted via an indicator variable for whether the calculated support interval contained $\hat{\theta}$. The coverage probability over B bootstrap repetitions is then calculated as

$$P(SI_{lower} < \theta < SI_{upper}) \approx \frac{\sum_{i=1}^{B} I_{\hat{\theta} \in SI_i}}{B}$$

where $[SI_{lower}, SI_{upper}]$ are random variables representing the random endpoints of the 1/k support interval. This process can be repeated at different k until nominal coverage is reached. One sample of B=10,000 bootstrap samples suggested a $1/7.245$ SI would be an approximate 95% CI for this data.

4.2.13 Smooth Bootstrap Calibrated Binomial

The binomial case merits special attention. As described in Brown, Kai, and DasGupta (25), the coverage probability of confidence procedures for binomial proportions demonstrates highly erratic behavior. The inherent problem is that the bootstrap procedure assumes that coverage probability at the true value, θ_0, is similar to the coverage probability at $\hat{\theta}$. However, this can hardly be expected from the coverage probabilities of CIs for binomial proportions. Given that values of θ which differ by miniscule amounts can have several percentage points difference in their actual coverage probability, doing a parametric bootstrap using $\hat{\theta}$ might lead to very poor estimate of the true coverage depending on the difference in coverage at θ versus $\hat{\theta}$, Figure 4.1.

If we were to attempt to probabilistically calibrate a support interval to demonstrate

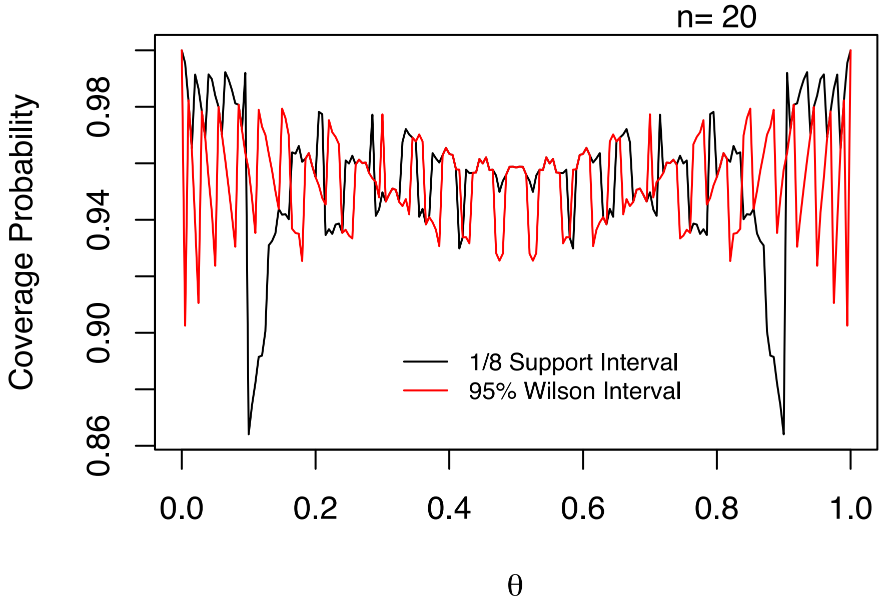


Figure 4.1: Coverage rates of Wilson Interval and 1/8 support intervals for a given true proportion and n=20.

the proper coverage rate using the previously described technique of treating k as a tuning parameter until nominal coverage is observed, the resulting relationship between k and coverage probability is a step function, Figure 4.2.

As such, it is not necessarily possible to find a value of k which gives exactly $1-\alpha$ coverage for a given sample size and true proportion, θ_0. A reasonable alternative would be to find the shortest interval, and therefore lowest k, that has at least $1-\alpha$ coverage. This is equivalent to finding the step in the k versus coverage function that steps over the $1-\alpha$ value.

For a fixed sample size we can numerically determine the proper k value given the value of the true proportion, Figure 4.3. The resulting function relating θ_0 to k_{95} is the key to probabilistic calibration. However, the resulting calibration will suffer from the same affliction that plagues all of the other confidence intervals for the binomial proportion.

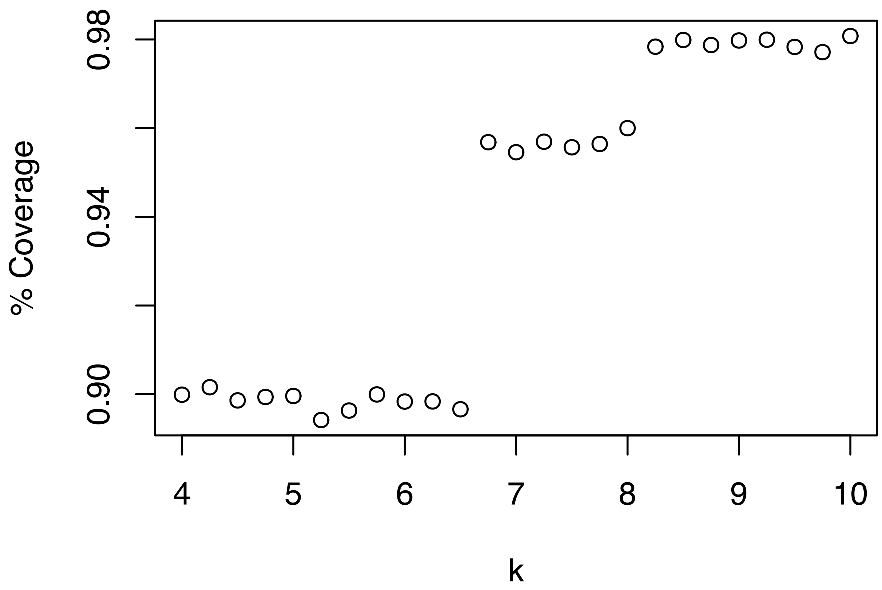


Figure 4.2: Coverage rate as a function of k with $\theta_0 = .2, n = 20$.

Specifically, proper calibration requires knowledge of the true value of θ_0 prior to estimating θ_0. For example, if an analyst wanted to estimate a support interval for $\theta_0 = 0.1$, then the appropriate choice of k_95 is approximately 8.23. However if they attempt to estimate the interval from data $x = 3, \hat{\theta} = 0.15$, they would perform a standard parametric bootstrap procedure and incorrectly conclude that the appropriate value of k_{95} necessary to achieve 95% coverage was approximately 11.49, i.e. they would evaluate the θ vs k_{95} function at the wrong location resulting in a longer than necessary interval that may or may not overcover.

Choosing the wrong k with which to calibrate the support interval can have a variety of consequences. In examining Figure 4.3, there is a notable discontinuity between $\theta_0 \in [0.31, 0.32]$. This region is a direct result of the step function relationship between k and the coverage probability demonstrated in Figure 4.2. This region of the θ space occurs when

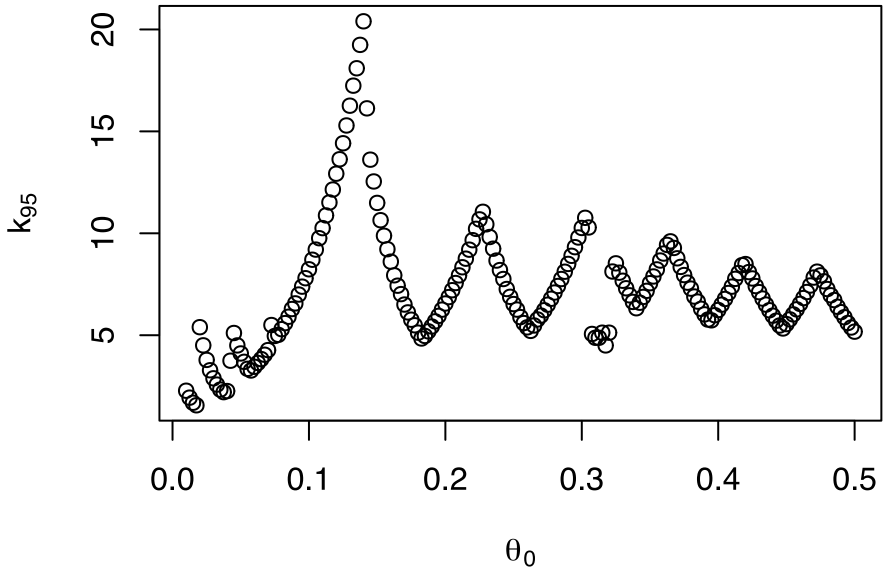


Figure 4.3: Value of *k* which results in 95% coverage given the true proportion, θ_0 for a sample of size 20.

one of the steps in the coverage function happens to coincide with the nominal coverage (the function suggests $(1-\alpha)*100\%$ coverage over a wide variety of k). Choosing the wrong *k* when θ_0 is truly in this range will likely maintain the nominal coverage. At other values of θ_0 the step function is significantly more volatile which can result in over/under coverage in addition to the consequences on interval length.

The proposed solution to this difficulty in selecting an appropriate *k* is to take a weighted average of *k* values over the range of plausible proportions. This is accomplished via a multilevel parametric bootstrap. At the first level, values for θ are chosen to be used in a second parametric bootstrap. They are chosen at random with a probability that is proportional to their likelihood, i.e. values are sampled from the posterior distribution of θ resulting from employing a uniform prior on θ. The second parametric bootstrap is then performed as before for successive levels, k, until nominal coverage is attained. As evident in Figure

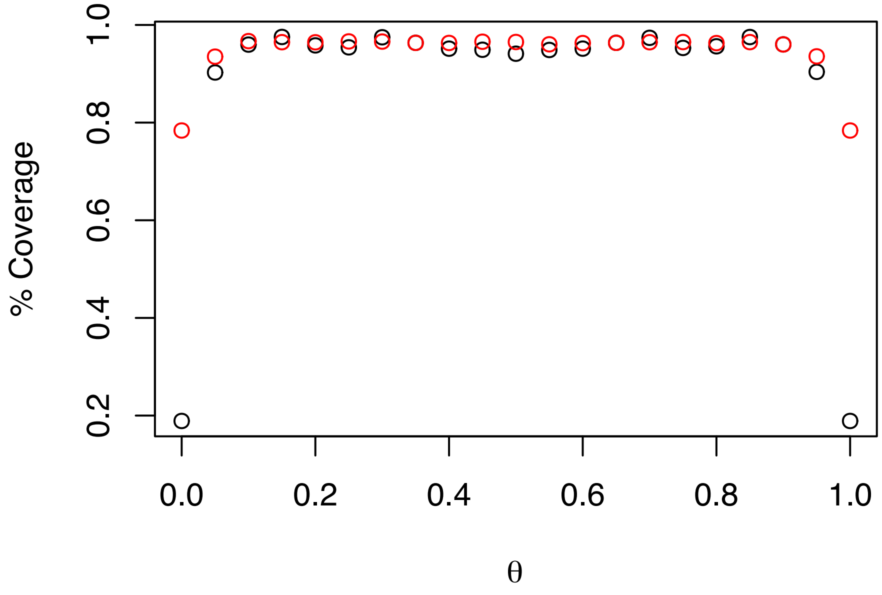


Figure 4.4: Estimated coverage given a particular observed value of $\hat{\theta}$ simulated from a uniform distribution of θ values.

4.4, the proposed method of smoothing demonstrates operating characteristics which are fairly consistent over the entire range of observable values of $\hat{\theta}$. In other words, an analyst who does not wish to leverage any prior knowledge about θ would expect more consistent performance of confidence intervals regardless of the observed data value. For the sake of comparison, the figure also shows the result of using the single level bootstrap that simply estimates the value of k at $\hat{\theta}$. The smoothed version shows particular promise in the upper and lower extremes of the θ scale.

4.2.14 Conclusion

The *supportInt* package provides tools for calculating likelihood based support intervals in R, which have been extremely limited at this point. The package utilizes a novel bootstrap technique to achieve probabilistic calibration of support intervals which narrows the gap between probability based and likelihood based inference. Although the ideological differences remain vast, it is now possible for the devout frequentist to enjoy $(1-\alpha)$ confidence intervals that maintain the favorable properties of support intervals. Similarly, the devout likelihoodist may report 1/k support intervals as 95% confidence intervals without fear of reprisal from reviewers bent on probability based inference.

REFERENCES

[1] Herbert Robbins. An Empirical Bayes Approach to Statistics. *Proceedings of the Third Berkeley Symp. on Math. Statist. and Prob.*, 1956.

[2] Charles Stein. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. *Proceedings of the Third Berkeley Symp. on Math. Statist. and Prob.*, 1956.

[3] Willard James and Charles Stein. Estimation with Quadratic Loss. *Proceedings of the Fourth Berkeley Symp. on Math. Statist. and Prob.*, 1963.

[4] Bradley Efron and Carl Morris. Stein's Estimation Rule and Its Competitors- An Empirical Bayes Approach. *J. Amer. Statist. Assoc.*, (341), 1973.

[5] Bradley Efron and Carl Morris. Data Analysis Using Stein's Estimator and its Generalizations. *J. Amer. Statist. Assoc.*, (350), 1975.

[6] Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data.* Springer, 2000.

[7] Nan Laird and James Ware. Random-Effects Models for Longitudinal Data. *Biometrics*, (4), 1982.

[8] Art Owen. *Empirical Likelihood.* CRC Press, 2001.

[9] B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 1979.

[10] C. V. Thakar. Perioperative acute kidney injury. *Advances in Chronic Kidney Disease*, 20:1, 2013.

[11] S. Calvert and A. Shaw. Perioperative acute kidney injury. *Perioperative Medicine*, 1:6, 2012.

[12] R. Bellomo, Ronco C., Kellum J.A., and Palevsky P. Mehta R.L. Acute renal failure-definition, outcome measures, animal models, fluid therapy and information technology needs: the second international consensus conference of the acute dialysis quality initiative (adqi) group. *Crit Care*, 2004.

[13] R. L. Mehta, J. A. Kellum, A.V. Shah, B. A. Molitoris, C. Ronco, D. G. Warnock, and A. Levin. Acute kidney injury network: report of an initiative to improve outcome in acute kidney injury. *Crit Care Med*, 11, 2007.

[14] Acute Kidney Injury Work Group. Kidney disease: Improving global outcomes (kdigo)- clinical practice guideline for acute kidney injury. *Kidney Inter*, 2012.

[15] R.L. Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*, 1989.

[16] A.P. Dempster and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J Royal Statist Soc B*, 1977.

[17] Karl W. Broman, Hao Wu, Saunak Sen, and Gary A. Churchill. R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19:889–890, 2003.

[18] B. Emma Huang and Andrew W. George. R/mpmap: a computational platform for the genetic analysis of recombinant inbred lines. *Bioinformatics*, 27:727–729, 2011.

[19] Richard Royall. *Statistical Evidence*. Chapman and Hall/CRC, 1997.

[20] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[21] Leena Choi. *ProfileLikelihood: Profile Likelihood for a Parameter in Commonly Used Statistical Models*, 2011. R package version 1.1.

[22] Brad Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, (2), 1988.

[23] Bradley Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, (1), 1979.

[24] Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statist. Sci.*, (3), 1996.

[25] Lawrence D. Brown, Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statist. Sci.*, (2), 2001.