Simulated Power Analysis of the Cox Proportional Hazards Model

Nicholas Vogt

Advisor: Keshav Pokhrel, Ph.D.

Department of Mathematics & Statistics
College of Arts, Sciences, & Letters
University of Michigan - Dearborn

December 2018

Abstract

The Cox proportional hazards model is a core component of several experimental designs in medicine and beyond. These experimental results impact the lives of millions of patients globally and billions of dollars in investment. Little is understood, however, regarding which design elements most contribute to a design's success or failure; i.e. the study's statistical power. Scientists can design more effective and cost-effective studies by understanding which design elements impact statistical power. We design an experiment to answer this question by simulating survival datasets, fitting several Cox proportional hazards models, and analyzing the results. Most notably, we find that short studies of several patients are more powerful than longer studies with fewer patients.

1 Introduction

The Cox proportional hazards model [2] is a core component of several experimental designs in medicine and beyond. These experimental results impact the lives of millions of patients globally and billions of dollars in investment. Little is understood, however, regarding which design elements most contribute to a design's success or failure; i.e. the study's statistical power. Scientists can design more effective and cost-effective studies by understanding which design elements impact statistical power.

1.1 An Illustrative Example

By way of example, consider the hypothetical osteopathic doctor, Caitlin. Caitlin runs a private practice and wants to verify that ibuprofen helps to break a fever. Flu season starts and she accepts 10 feverish patients daily. Each consenting patient is prescribed ibuprofen and reports their body temperature back to Caitlin in the evening.

After the first day, Caitlin understands that she probably doesn't have enough data to find an effect; her data lacks statistical power. Almost every patient is still feverish because not enough time has passed for the ibuprofen to break the fever.

On the 10^{th} day, Caitlin thinks she has enough data to test her hypothesis. She wonders, however, if she is correct to remove the 10^{th} day's data from her analysis. For the same reason she did not model the first day's patients for lack of statistical power, so too should she omit the patients on the 10^{th} day? And if the 10^{th} day is invalid, does she remove days nine and eight? She doesn't know the answer, but she's confident she should filter out day 10 at a minimum.

One solution is to stop introducing new patients to the study and to analyze the results after every existing patient reports that their fever is gone. This ensures an uncensored dataset and delays her study indefinitely. If ibuprofen is a fever reducer, then Caitlin misses the opportunity to treat new patients more effectively. While the cost of waiting is small when treating a fever, a more aggressive ailment is more costly.

A second solution is to observe the data and pick the cut-off point that looks best, but Caitlin is wary of biasing the dataset by choosing data she determines valuable. Her results may be called into question and compromise her integrity as a medical professional.

How then, can Caitlin honestly and ethically subset her dataset to maximize the power of the Cox proportional hazards model? The answer is important, not only to Caitlin, but to medicine in general. If we can design a better study, then we can test results and apply the resulting treatments to patients faster.

1.2 Existing Literature

There is some existing literature on the topic of power analysis for the Cox proportional hazards model [1, 3, 5], but none recommend how to filter a dataset to maximize statistical power of the Cox proportional hazards model. This is a novel approach as far as we can tell.

2 Methodology

We define and outline the methodology to estimate the power of given arguments to a Cox proportional hazards model. First, we define simulation parameters to construct the simulated datasets in R. We then analyze the resulting parameter estimates of the Cox proportional hazards model on each generated dataset. Simulation parameters that correctly identify a significant effect are

labeled "correct" or "incorrect." We merge the original simulation parameters and the correctness of the Cox proportional hazards model, and evaluate logistic regression on the resulting dataset. The result of this logistic regression is our power estimate.

2.1 Simulated Datasets

We construct a dataset from six parameters[1],

- baseline hazard rate (λ)
- treatment hazard ratio ($\exp[\beta]$)
- periods observed (n_p)
- sequential cohorts observed (n_c)
- cohort size (s)
- random seed (ξ).

The resulting dataset contains $n_c \times s$ observations on simulated individuals. The binomial treatment variable is evenly distributed across individuals such that $\frac{s}{2}$ are labeled "treated" and the remaining are labeled "untreated".

We must classify whether an individual died or if they are right-censored. Individual time of death is sampled from the negative binomial distribution[2]

$$NB(1, \lambda \cdot \exp[\beta x]) \tag{1}$$

where

$$x = \begin{cases} 1, & \text{if treated} \\ 0, & \text{otherwise} \end{cases}$$

The observation is right-censored where equation (1) is greater than the max periods observed on that individual.

By example, the parameters $\lambda = 0.5,\ \exp[\beta] = 0.2,\ n_p = 4,\ n_c = 2,\ s = 2$ may yield the dataset

PeriodAdmitted	IsTreated	HazardRate	IsDeceased	LastObserved
1	0	0.5	1	1
1	1	0.1	0	4
2	0	0.5	1	2
2	1	0.1	1	3

[1] The dataset simulations was programmed with R in RStudio [7, 8]. The source code can be found on the author's GitHub, vogt4nick/coxph-power-analysis [10].

[2] We define the hazard rate as $\lambda \cdot \exp[\beta x]$ to reflect Cox's proportional hazards assumption.

2.1.1 Choosing simulation parameters

A common problem encountered by researchers is how to determine which model to use. We find a rare reversal: How do we choose which datasets to use? More specifically, how do we choose which parameters to model? There are three concerns.

The first concern is to choose simulation parameters which will generate datasets with values such that the Cox proportional hazards model converges. The results are invalid otherwise.

The second concern is to choose simulation parameters which are reasonably uniformly distributed in the logit space. The inverse logit function and powers of 2 are good candidates for this. We set the simulation parameters as follows

- $\lambda \in \{logit^{-1}(i) \mid \forall i \in \{-6, -5, -4, -3, -2, -1, 0\}\}$,
- $\beta \in \{logit^{-1}(i) \mid \forall i \in \{-3, -2, -1, 0, 1, 2\}\}$,
- $n_p \in \{2^i \mid \forall i \in \{2, 3, 4, 5, 6\}\}$,
- $n_c \in \{2^i \mid \forall i \in \{2, 3, 4, 5, 6\}\}$,
- $s \in \{2^i \mid \forall i \in \{2, 3, 4, 5\}\}$,
- $\xi \in \{1, 2, 3, ..., 19, 20\}$

and create datasets for every combination of parameters.

The third and final concern is how to choose simulation parameters that will create datasets whose observed power is not only 0 or 1. We view this as a sampling problem. We calculate the observed power, $\hat{\rho}$, and sample five observations at each level in $(\frac{0}{20}, \frac{1}{20}, \frac{2}{20}, ..., \frac{20}{20})$. This methodology uniformly balances the distribution of the target variable at the cost of reducing the size of the final dataset.

2.1.2 Cox Proportional Hazards Model

The Cox proportional hazards model is specified as

$$h(t) = \theta \exp[\alpha x] + \varepsilon \tag{2}$$

The key to Cox's formulation, and hence the model's namesake, is the decision to model the hazard rate, $h(t)$, as the proportional effect of a treatment variable, α, on the baseline hazard rate, θ. The regression estimates θ and α correspond to the simulation parameters λ and β respectively. We later compare the estimates to the simulation parameters to determine the efficacy of the fitted model.

2.2 Define Features

We identify three features to help fit the logistic regression model. Two are predictors which are more informative than the basic simulation parameters.

The third feature determines the "correctness" of the model, and will define the target variable for logistic regression.

The length of the study is best framed as a function of the baseline hazard rate. Rather than counting observed periods, we define

$$\nu = \frac{n_p}{\log_{1-\lambda} \frac{1}{2}} \tag{3}$$

as the number of periods an individual with hazard rate λ is expected to survive. Note that $\log_{1-\lambda} \frac{1}{2}$ is the expected value of $NB(1, \lambda)$.

Next, rather than set n_p and n_c as model parameters, we define

$$\omega = \frac{n_c}{n_p} \tag{4}$$

as the fraction of enrollment periods that accept new cohorts.

2.2.1 Correctness

We intend to use logistic regression to identify which dataset parameters have the greatest effect on statistical power; which dataset parameters maximize the chance of a correct result. Before continuing further, we define what constitutes a "correct" result. Classically, our null hypothesis is

$$H_0 : \alpha = 0 \tag{5}$$

Since we also defined the true α (i.e. β) to be always less than 0, we define another hypothesis

$$H_1 : \alpha < 0 \tag{6}$$

The estimated model is labeled "correct" if we reject hypotheses H_0 and H_1. That is,

$$\pi = \begin{cases} 1, & \text{if } \neg H_0 \wedge \neg H_1 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

We use a Wald test [11] to determine whether a result rejects a hypothesis. That is, we reject H_0 if

$$\frac{|\hat{\alpha} - 0|}{\text{se}(\hat{\alpha})} < z_{0.975}$$

In the same fashion, we reject[3] H_1 if

$$\frac{\hat{\alpha} - 0}{\text{se}(\hat{\alpha})} < -z_{0.975}$$

[3]It's worth observing that π is equivalent to $\neg H_0 \wedge \alpha < 0$ in our special case.

2.3 Predict Power with Logistic Regression

Now, with the target variable identified, we can evaluate logit regression on the dataset. We propose two models for predicting the statistical power of the Cox proportional hazards model.

The first models linear effects

$$\text{logit}(\pi) = \gamma_0 + \gamma_1\lambda + \gamma_2\beta + \gamma_3\nu + \gamma_4\omega + \gamma_5 s \tag{8}$$

The linear-effects model ought to confirm what we believe about the relationship between the simulation parameters and π. We'll pay particular attention to the sign of the effect.

The second model examines quadratic effects,

$$\begin{aligned}\text{logit}(\pi) =& \gamma'_0 + \\ & \gamma'_{11}\lambda + \gamma'_{12}\lambda^2 + \\ & \gamma'_{21}\beta + \gamma'_{22}\beta^2 + \\ & \gamma'_{31}\nu + \gamma'_{32}\nu^2 + \\ & \gamma'_{41}\omega + \gamma'_{42}\omega^2 + \\ & \gamma'_{51}s + \gamma'_{52}s^2\end{aligned} \tag{9}$$

Building on the linear-effects model, the quadratic effects model will help us better understand the relationship between the simulation parameters and π. Specifically, we'll look for strong effect sizes and opposite signs on the x and x^2 terms. Opposite signs will indicate an inflection point at some $x > 0$.

The cubic-effects model builds further on the quadratic-effects model,

$$\begin{aligned}\text{logit}(\pi) =& \gamma^{(2)}_0 + \\ & \gamma^{(2)}_{11}\lambda + \gamma^{(2)}_{12}\lambda^2 + \gamma^{(2)}_{13}\lambda^3 + \\ & \gamma^{(2)}_{21}\beta + \gamma^{(2)}_{22}\beta^2 + \gamma^{(2)}_{23}\beta^3 + \\ & \gamma^{(2)}_{31}\nu + \gamma^{(2)}_{32}\nu^2 + \gamma^{(2)}_{33}\nu^3 + \\ & \gamma^{(2)}_{41}\omega + \gamma^{(2)}_{42}\omega^2 + \gamma^{(2)}_{43}\omega^3 + \\ & \gamma^{(2)}_{51}s + \gamma^{(2)}_{52}s^2 + \gamma^{(2)}_{53}s^3\end{aligned} \tag{10}$$

The role of the cubic-effects model is primarily to satiate our curiosity, and may only add to our inference.

The fourth and final model examines diminishing effects,

$$\text{logit}(\pi) = \gamma^{(3)}_0 + \gamma^{(3)}_1 \exp[\lambda] + \gamma^{(3)}_2 \exp[\exp[\beta]] + \gamma^{(3)}_3 \ln(\nu) + \gamma^{(3)}_4 \ln(\omega) + \gamma^{(3)}_5 \ln(s) \tag{11}$$

Note that the diminishing effects model calculates $\exp[\lambda]$ and $\exp[\exp[\beta]]$. We expect a diminishing, inverse relationship. That is, we anticipate a diminishing effect as λ and $\exp[\beta]$ get smaller.

We use the resulting fitted models to predict the statistical power of a Cox proportional hazards model applied to the dataset. To further evaluate the appropriateness of each model, we compare the Akaike information criterion (AIC) [6, 9] for each model.

3 Results

We analyze the results to understand which dataset features most affect the power of a Cox proportional hazards model. We evaluate the value of the simulated datasets under our specifications, and build a model to estimate the power of the Cox proportional hazards model under our specifications.

3.1 Quality of Simulated Data

We acknowledged three problems in choosing appropriate dataset parameters:

1. How to choose dataset parameters which will generate datasets with values such that the Cox proportional hazards model converges?
2. How to choose dataset parameters which are uniformly distributed in the logit space?
3. How to choose dataset parameters to simulate datasets whose observed power, $\hat{\rho}$, is uniformly distributed between 0 and 1?

We verify that chose appropriate dataset parameters by observing the distribution of observed power (Figure 1). The mode power is 0, however the values between 0 and 1 are reasonably uniform. Our decision to resample 5 observations from each level holds.

3.2 Logistic Models

In this section we interpret the predicted effects of four models. We briefly review the model results before discussing the effects of dataset parameters separately. The estimates for each model are summarized in tables 1 and 2 [4].

We begin with the linear-effects model which notably finds a startling, positive estimate on the effect of γ on π. That said, the result is not reflected in any of the other models. Often, a result such as this can be attributed to multicollinearity, or a local, non-linear trend dominating the fit of the model, however, we took certain precautions to avoid such behavior in our model (Section 2.1.1). An explanation evades us at the time of writing. The remaining coefficients on the linear-effects model confirm our prior hypotheses and are otherwise unremarkable.

The quadratic-effects model uncovered more powerful effects on each simulation parameter except λ. Furthermore, the signs on all estimates take the form $-x^2 + x$, indicating an inversion point at some $x > 0$. These estimates may come to support the diminishing effects model. Finally, the AIC of the

quadratic effects model is smaller than the linear-effects model, indicating a more appropriate model (2,445 and 2,651 respectively).

The cubic-effects model adds little evidence the quadratic-effects model does not already present. Nonetheless, the AIC indicates a subtle improvement upon the quadratic-effects model (2,394 and 2,445 respectively). We will further investigate the individual effects in the next section.

Finally, we examine the diminishing-effects model. Every estimate has the expected sign, indicates a strong effect size, and each is significant at the $p < 0.01$ level. The AIC is similar to that of the quadratic-effects model (2,430 and 2,445 respectively). We prefer the diminishing-effects model for its interpretive power, as it removes much of the complexity of the quadratic- and cubic-effects models.

3.2.1 Baseline Hazard Rate (λ)

Intuitively, there is a strictly inverse relationship between λ and π; a study is more likely to find evidence for a true effect if the observation window is small. In accordance with this belief, all but the linear-effects model agrees with this assertion; a smaller λ increases power (Figure 2). The effect is not as pronounced as we suspected. For example, reducing the observation window by a factor of two, and therefore reducing λ by half, adds little power to the study.

3.2.2 Treatment Hazard Ratio ($\exp[\beta]$)

A larger $\exp[\beta]$ implies a smaller treatment effect[4], and smaller treatment effect is be harder to identify. The linear- and diminishing-effects model behave as anticipated (Figure 3). The diminishing-effects model estimates a sigmoid-like behavior with a larger marginal effect for values $\exp[\beta] \in [0.4, 0.6]$.

The quadratic- and cubic-effects models appear to have overfit the trend identified by the diminishing-effects model; the marginal effect inverts from negative to positive. We prefer this explanation, however, we posit another unverified explanation: if the treatment all but guarantees immunity, then the treated individuals in our simulation would all be right-censored. This prompts a follow up question: Is the Cox proportional hazards model robust to unbalanced censored data?

3.2.3 Cohort Size (s)

Figure 4 shows the estimated relationship of cohort size on power for each model.

The anticipated, positive relationship of s on pi is substantiated by all four models (Figure 4). Of all the simulation parameters, this one seems best explained by the log-effects model. The marginal effect of the fifth patient is intuitively larger than the marginal effect of the 25^{th} patient. Both the quadratic- and cubic-effects models show a second-order inversion which agrees with the diminishing effect assumption.

[4]Recall we defined values for β such that $\exp[\beta] < 1$.

3.2.4 Expected Lifetimes (ν)

Expected lifetimes (ν) proxies study-length. Power rapidly increases where $\nu < 1$ and slowly tapers off after $\nu = 2$. The effect is so pronounced that it seems reasonable to advise small studies observe at least one expected lifetime before applying the Cox proportional hazards model, when possible. However, this effect is readily offset by the introduction of large cohorts (Figure 6). Every model supports the notion that short studies of several patients are more powerful than longer studies with few patients.

3.2.5 Share of Open Enrollment Periods (ω)

Finally, we arrive at the core question we set out to answer. How do we filter our dataset to maximize power? The quadratic- and diminishing-effects models disagree where $\omega > 0.8$ (Figure 7). The former suggests including the last 20% of cohorts cause the study to lose power power, whereas the latter suggests the same add power to the study. With these results, we believe the answer is unimportant; the effect of the final 20% of cohorts is very small in either case[5].

4 Discussion

We set out to understand how a scientist can ethically subset their dataset to maximize the stastical power of the Cox proportional hazards model. We designed our experiments to understand how different design elements affect a study's power.

We close with three findings:

1. Reducing the observation period adds comparatively little power to the study.
2. Short studies of several patients are more powerful than longer studies with fewer patients.
3. The final 20% of cohorts present very little value to the statistical power of the study.

With these results, we recommend studies divert resources to short studies of several patients where possible. Researchers can further reduce cost by closing enrollment for the final 20% of observation periods.

4.1 Next Steps

We present recommendations for future researchers interested in reproducing and expanding on our work. The most immediate improvements can be made by defining new features and fitting new models on these features. In the absence

[5]It's likely there is an interaction effect between ν and ω. Naturally, a longer study should be able to include more cohorts.

of a theoretically derived power calculation, a reliable predictive model may be employed with sufficient data and model design. Such a predictive model could inform scientists globally and reduce the cost of research. Finally, we'd like to see our findings applied to different survival models, such as the accelerated failure time model to compare and contrast results.

References

[1] Bender, R., Augustin, T., and Blettner, M. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine 24*, 11 (2005), 1713–1723.

[2] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34:2* (1972), 187–220.

[3] Gönen, M., and Heller, G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika 92*, 4 (2005), 965–970.

[4] Hlavac, M. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Central European Labour Studies Institute (CELSI), Bratislava, Slovakia, 2018. R package version 5.2.2.

[5] Hsieh, F., and Lavori, P. W. Sample-size calculations for the cox proportional hazards regression model with nonbinary covariates. *Controlled clinical trials 21*, 6 (2000), 552–560.

[6] Hyndman, R. Facts and fallacies of the aic. https://robjhyndman.com/hyndsight/aic/.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[8] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015.

[9] Sakamoto, Y., and Kitagawa, G. *Akaike Information Criterion Statistics*. Kluwer Academic Publishers, Norwell, MA, USA, 1987.

[10] Vogt, N. Cox proportional hazards power analysis. https://github.com/vogt4nick/coxph-power-analysis, 2018.

[11] Wald, A. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics 16:2*.

Tables

Figures

Table 1: Linear-, Quadratic-, & Cubic-Effects Models of Statistical Power

	Dependent variable:		
	π		
	(Linear Effects)	(Quadratic Effects)	(Cubic Effects)
λ	1.340^{***}	0.751	-2.235
	(0.393)	(1.957)	(4.326)
λ^2		-4.985	-22.803
		(3.371)	(22.338)
λ^3			43.008
			(30.519)
$\exp(\beta)$	-2.578^{***}	4.912^{***}	11.310^{***}
	(0.168)	(0.719)	(2.057)
$\exp(\beta)^2$		-9.106^{***}	-28.417^{***}
		(0.848)	(5.509)
$\exp(\beta)^3$			14.215^{***}
			(3.981)
s	0.011^{**}	0.100^{***}	0.247^{*}
	(0.005)	(0.025)	(0.132)
s^2		-0.002^{***}	-0.014
		(0.001)	(0.010)
s^3			0.0003
			(0.0002)
ν	0.003	0.171^{***}	0.439^{***}
	(0.006)	(0.022)	(0.051)
ν^2		-0.002^{***}	-0.016^{***}
		(0.0003)	(0.002)
ν^3			0.0001^{***}
			(0.00002)
ω	-0.022	1.830^{**}	12.581^{***}
	(0.149)	(0.718)	(3.011)
ω^2		-1.147^{*}	-25.660^{***}
		(0.598)	(7.291)
ω^3			15.018^{***}
			(4.628)
γ_0	0.823^{***}	-1.327^{***}	-3.402^{***}
	(0.119)	(0.252)	(0.577)
Observations	2,098	2,098	2,098
Log Likelihood	$-1,319$	$-1,212$	$-1,181$
Akaike Inf. Crit.	2,651	2,445	2,394
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 2: Diminishing-Effects Model of Statistical Power

	Dependent variable:
	π
	(Diminishing Effects)
$\exp(\lambda)$	-2.391^{***}
	(0.356)
$\exp(\exp(\beta))$	-2.550^{***}
	(0.141)
$\ln(s)$	0.748^{***}
	(0.084)
$\ln(\nu)$	0.806^{***}
	(0.064)
$\ln(\omega)$	0.432^{***}
	(0.066)
γ_0	5.228^{***}
	(0.484)
Observations	2,098
Log Likelihood	$-1,209$
Akaike Inf. Crit.	2,430
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 1:

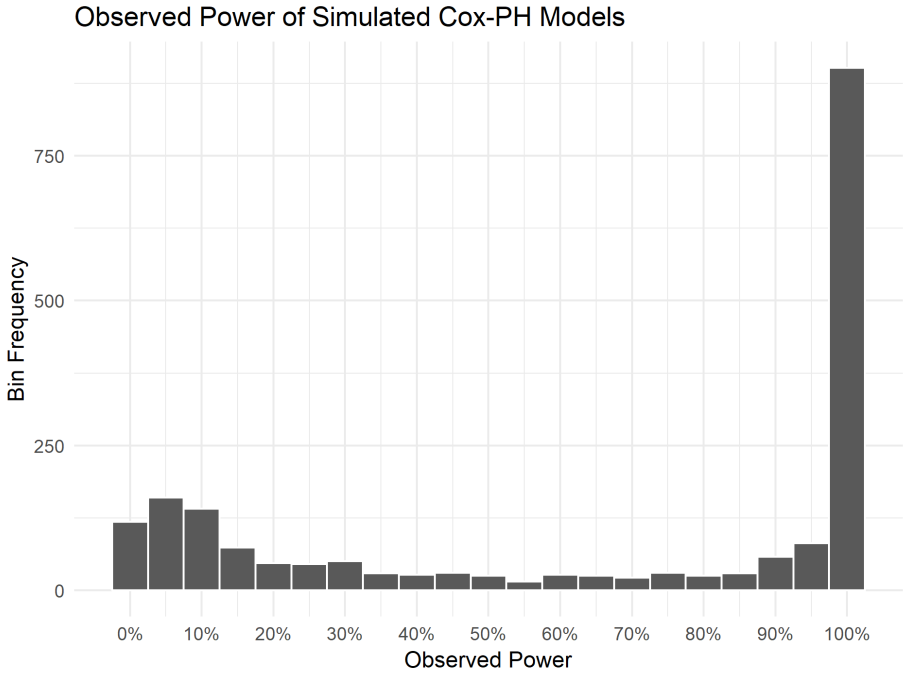


Figure 2:

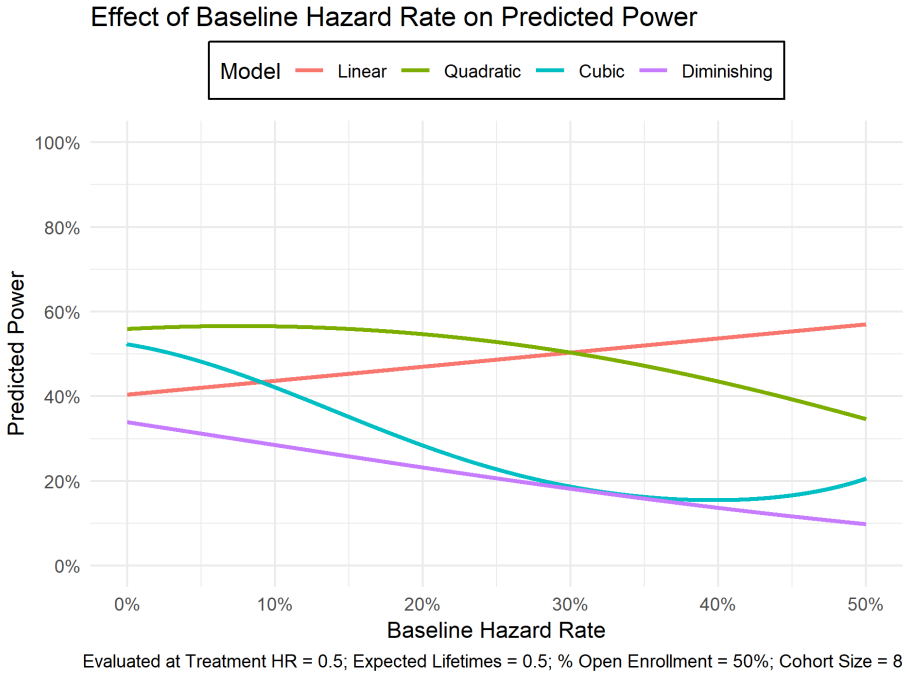


Figure 3:

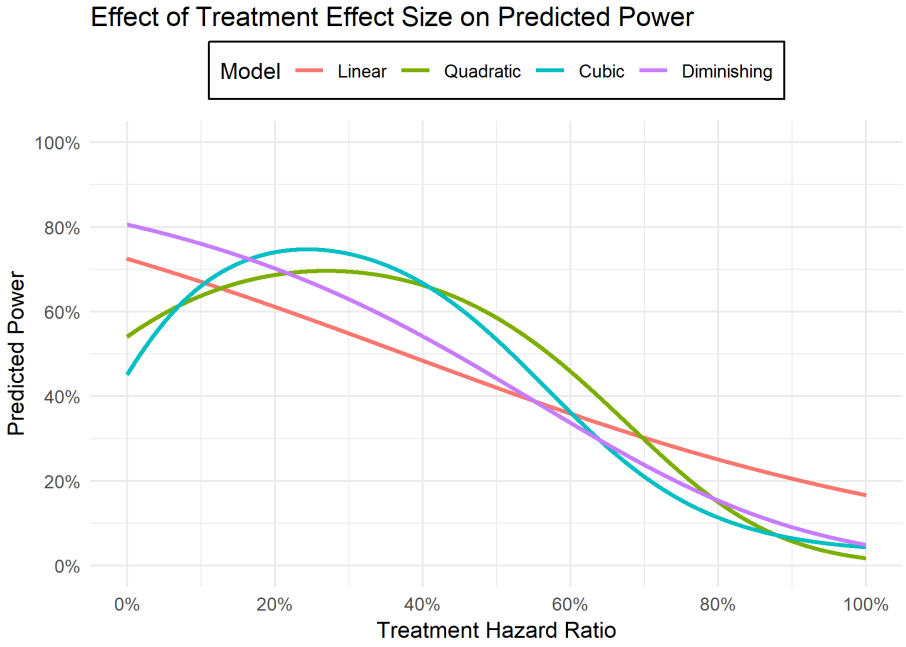


Figure 4:

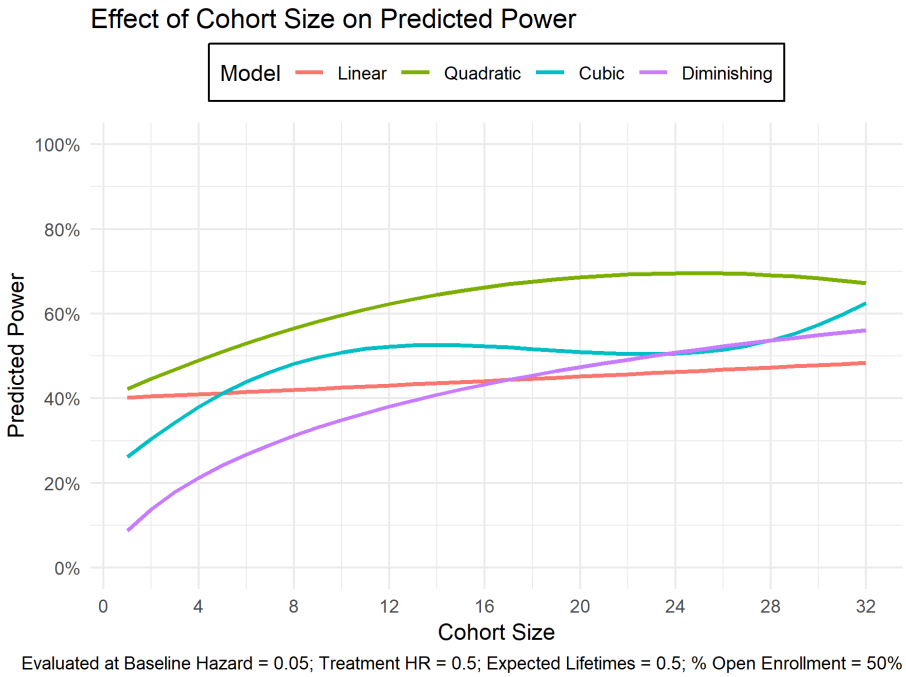


Figure 5:

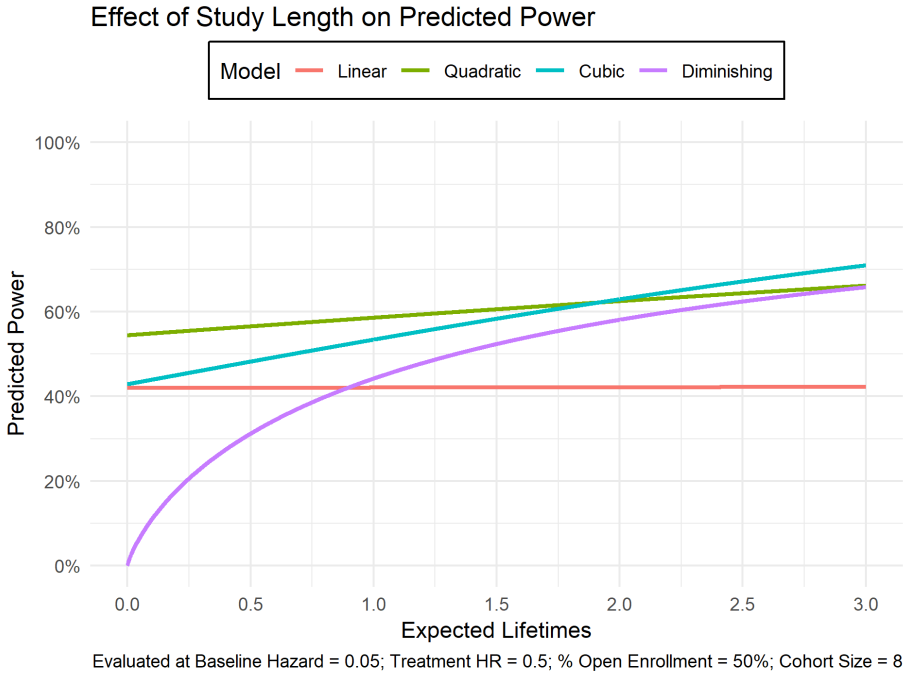


Figure 6:

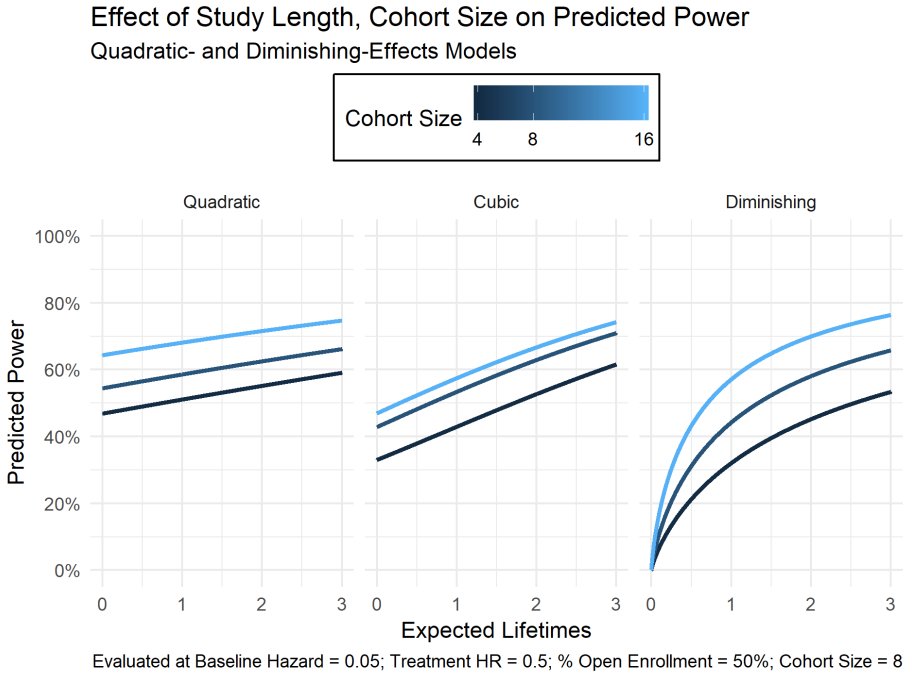
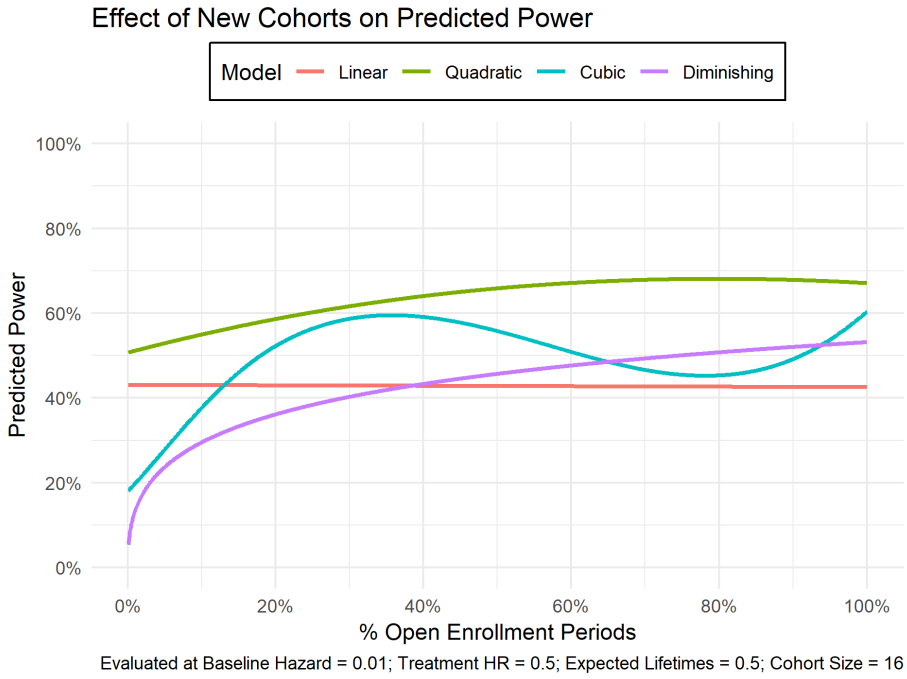


Figure 7:


Effect of New Cohorts on Predicted Power
Model
Linear
Quadratic
Cubic
Diminishing
Predicted Power
100%
80%
60%
40%
20%
0%
0%
20%
40%
60%
80%
100%
% Open Enrollment Periods
Evaluated at Baseline Hazard = 0.01; Treatment HR = 0.5; Expected Lifetimes = 0.5; Cohort Size = 16