Methods for Hypothesis Testing in Animal Carcinogenicity Experiments

Anna Korpak

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Barbara McKnight, Chair

Brian G. Leroux

James P. Hughes

Program Authorized to Offer Degree:
Public Health - Biostatistics

University of Washington

Abstract

Methods for Hypothesis Testing in Animal Carcinogenicity Experiments

Anna Korpak

Chair of the Supervisory Committee:
PhD Barbara McKnight
Biostatistics

Animal carcinogenicity experiments are conducted by private entities and government agencies to investigate whether a substance causes cancer. Since most tumors are occult and it is necessary to conduct a necropsy to detect the presence of cancer, producing unbiased tests for differences in tumor incidence is challenging. The highest doses given in these experiments often have toxic effects in the animals or might affect tumor lethality, and many common statistical methods for survival data perform poorly when such effects induce differences in mortality across treatment groups. The poly-k test, developed by Bailer & Portier (1988) as a modification of the Cochran-Armitage test, was designed to avoid problems from differential mortality, but is based on strong parametric assumptions, specifically that underlying tumor hazard may be modeled as Weibull with shape parameter k (usually set to 3). Existing literature that examines the performance of the poly-k and competitor tests assumes Weibull tumor hazards, and finds that the poly-k can be biased under treatment toxicity when its shape parameter assumption is not met. Given that this test has become a standard for government agencies such as the National Toxicology Program and the FDA, closer examination is warranted. Our simulations examine the performance of the poly-k under non-Weibull tumor hazards, and find that our own parametric tests can outperform it under these conditions. For tests like the poly-k, our goal is to develop and examine the

performance of adaptive testing algorithms that estimate a test's parameter based on the data; one such estimation was suggested under Weibull assumptions by Moon et al. (2003), though their approach based on estimating lifetime cumulative tumor incidence rates does not make full use of the available data, and requires that the experiment include multiple interim sacrifices for adequate performance. Under most of our simulation settings, our poly-$\hat{k}_{MLE}$ test, based on estimating the shape parameter k by maximizing the full likelihood, has type I error and power very similar to the poly-k test with correctly-specified k, and it maintains size better, with comparable or higher power, than the test based on the Moon et al. estimate. Our MLE-based test does not require interim sacrifices, although it may perform better under some interim sacrifice experimental designs. We compare these tests under a variety of parametric assumptions and serial sacrifice designs to examine which experimental settings are most optimal for test performance.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables vi

Glossary viii

Chapter 1: Introduction 1
1.1 Carcinogenicity Experiments 1
1.2 Existing Methods 2
1.3 Literature on the poly-k test 4
1.4 Research Summary 7

Chapter 2: Effect of Violating the Weibull Assumption on the Performance of the Poly-3 Trend Test 9
2.1 Introduction 9
2.2 Alternate Cochran-Armitage-Based Tests 11
2.3 Simulation Study 14
2.4 Data Set Example 17
2.5 Results 19
2.6 Discussion 30

Chapter 3: An MLE-based estimation of k for the poly-k trend test 33
3.1 Introduction 33
3.2 Simulation Study 39
3.3 Results 42
3.4 Discussion 67

Chapter 4: Serial Sacrifices in Carcinogenicity Experiments 69
4.1 Introduction 69
4.2 Simulation Study 71
4.3 Results 73
4.4 Discussion 87

Chapter 5: Conclusions and Future Work 90
5.1 Conclusions 90
5.2 Future Work 93

Bibliography 95

Appendix A: Supplemental Materials for Chapter 1 99
A.1 Bieler & Williams Variance Calculation 99

Appendix B: Supplemental Materials for Chapter 2 104
B.1 Simulation Settings 104
B.2 Uncalibrated Power 106
B.3 Additional Simulation Results: 1-sided tests 108
B.4 Additional Simulation Results: 2-sided tests 110

Appendix C: Supplemental Materials for Chapter 3 114
C.1 Simulation Settings: Examining the shape of data 114
C.2 Simulation Output: bias & variability 120
C.3 Simulation Output: 1-sided tests 127
C.4 Simulation Output: 2-sided tests 133

Appendix D: Supplemental Materials for Chapter 4 139
D.1 Simulation Output: mean, SD, and MSE of $\widehat{k_T}$, comparing Moon and MLE estimation methods 139

LIST OF FIGURES

Figure Number		Page
2.1	Comparison of Hazards	12
2.2	Animal-specific time-at-risk quantities α_{ij} for different hazard assumptions	12
2.3	Data simulated under H_0 (type I error) with $p_0 = 0.15$, for nominal 0.05 level 1-sided tests	22
2.4	Data simulated under H_0 (type I error) with $p_0 = 0.15$, for nominal 0.05 level 1-sided tests (less extreme departures from poly-3 assumptions)	23
2.5	Data simulated under H_a (Power) with $p_0 = 0.15$, $p_1 = 0.3$, for 1-sided calibrated 0.05 tests	24
2.6	Data simulated under H_a (Power) with $p_0 = 0.15$, $p_1 = 0.3$, for 1-sided calibrated 0.05 tests (less extreme departures from poly-3 assumptions)	25
2.7	Comparison of CA-based tests with other competitors, by lethality	27
2.8	Data Set Example: Pulegone Cumulative Incidence Curves	28
3.1	Tumor hazards used in simulation studies	39
3.2	Mean $\hat{k}_T$ when H_0 true and $p_0 = 0.15$; varying k_T, lethality, toxicity (Weibull tumor hazard)	45
3.3	Mean $\hat{k}_T$ when H_a true and $p_0 = 0.15$; varying k_T, lethality, toxicity (Weibull tumor hazard)	46
3.4	Mean $\hat{k}_T$ when H_0 true and $p_0 = 0.05$; varying k_T, lethality, toxicity (Weibull tumor hazard)	47
3.5	Mean $\hat{k}_T$ when H_a true and $p_0 = 0.05$; varying k_T, lethality, toxicity (Weibull tumor hazard)	48
3.6	MSE of $\hat{k}_T$ when H_0 true and $p_0 = 0.15$; varying k_T, lethality, toxicity (Weibull tumor hazard)	49
3.7	MSE of $\hat{k}_T$ when H_a true and $p_0 = 0.15$; varying k_T, lethality, toxicity (Weibull tumor hazard)	50

3.8 MSE of $\hat{k}_T$ when H_0 true and $p_0 = 0.05$; varying k_T, lethality, toxicity (Weibull tumor hazard) 51
3.9 MSE of $\hat{k}_T$ when H_a true and $p_0 = 0.05$; varying k_T, lethality, toxicity (Weibull tumor hazard) 52
3.10 True tumor hazard (solid lines) for data generation versus Weibull with k parameter estimated as an average over many data sets 54
3.11 Type I error: $p_0 = 0.15$, varying k_T, lethality, toxicity (Weibull tumor hazard) 56
3.12 Power: $p_0 = 0.15$, varying k_T, lethality, toxicity (Weibull tumor hazard) . . . 57
3.13 Type I error: $p_0 = 0.05$, varying k_T, lethality, toxicity (Weibull tumor hazard) 58
3.14 Power: $p_0 = 0.05$, varying k_T, lethality, toxicity (Weibull tumor hazard) . . . 59
3.15 Type I error: $p_0 = 0.15$, varying k_T, lethality, toxicity (log-logistic tumor hazard) 60
3.16 Power: $p_0 = 0.15$, varying k_T, lethality, toxicity (log-logistic tumor hazard) . 61
3.17 Type I error: $p_0 = 0.15$, varying k_T, lethality, toxicity (Gompertz tumor hazard) 62
3.18 Power: $p_0 = 0.15$, varying k_T, lethality, toxicity (Gompertz tumor hazard) . . 63

4.1 Mean $\hat{k}$ by sacrifice strategy when H_0 true, varying lethality, toxicity, and k_T (note that n total (interim) sacrifices is 0 when number of sacrifice times = 1) 77
4.2 Mean $\hat{k}$ by sacrifice strategy when H_a true, varying lethality, toxicity, and k_T (note that n total (interim) sacrifices is 0 when number of sacrifice times = 1) 78
4.3 MSE of $\hat{k}$ by sacrifice strategy when H_0 true, varying lethality, toxicity, and k_T (note that n total (interim) sacrifices is 0 when number of sacrifice times = 1) 79
4.4 MSE of $\hat{k}$ by sacrifice strategy when H_a true, varying lethality, toxicity, and k_T (note that n total (interim) sacrifices is 0 when number of sacrifice times = 1) 80
4.5 Type I error: $p_0 = 0.15$, $L = 1$, $TOX = 1.8$, varying k_T 81
4.6 Type I error: $p_0 = 0.15$, $L = 1.5$, $TOX = 1$, varying k_T 82
4.7 Type I error: $p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$, varying k_T 83
4.8 Power: $p_0 = 0.15$, $L = 1$, $TOX = 1.8$, varying k_T 84
4.9 Power: $p_0 = 0.15$, $L = 1.5$, $TOX = 1$, varying k_T 85
4.10 Power: $p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$, varying k_T 86

B.1 Data simulated under Ha (un-calibrated Power) with varying p_0 (rows), effect size (line type), and true k_T (columns), for 1-sided tests 107
B.2 Data simulated under Ho (type I error) with $p_0 = 0.30$, for nominal 0.05 level 1-sided tests 108
B.3 Data simulated under Ho (type I error) with $p_0 = 0.05$, for nominal 0.05 level 1-sided tests 109
B.4 Data simulated under Ho (type I error) with $p_0 = 0.15$, for nominal 0.05 level 2-sided tests 110
B.5 Data simulated under Ha (Power) with $p_0 = 0.15$, $p_1 = 0.3$, for 2-sided calibrated 0.05 tests 111
B.6 Data simulated under Ho (type I error) with $p_0 = 0.15$, for nominal 0.05 level 2-sided tests (less extreme departures from poly-3 assumptions) 112
B.7 Data simulated under Ha (Power) with $p_0 = 0.15$, $p_1 = 0.3$, for 2-sided calibrated 0.05 tests (less extreme departures from poly-3 assumptions) 113

C.1 Simulations under Ho and Ha, with k=1, p_0=0.15, L=1, TOX=1.8, totsacr=10 116
C.2 Simulations under Ho and Ha, with k=1.5, p_0=0.15, L=1, TOX=1.8, totsacr=10 117
C.3 Simulations under Ho and Ha, with k=3, p_0=0.15, L=1, TOX=1.8, totsacr=10 118
C.4 Simulations under Ho and Ha, with k=6, p_0=0.15, L=1, TOX=1.8, totsacr=10 119

LIST OF TABLES

Table Number Page

2.1 Liver Cancer vs. Treatment Group 29
2.2 Bladder Cancer vs. Treatment Group 29
3.1 Simulation Experimental Design Settings: Serial Sacrifices 41
4.1 Serial Sacrifice Design: Varying Group Size and Total Interim Sacrificed . . . 72
B.1 Simulation Settings with Weibull, log-logistic, and Gompertz tumor hazards 105
C.1 Bias of $\hat{k}_T$ ($p_0 = 0.15$, $L = 1$, $TOX = 1$) 121
C.2 Bias of $\hat{k}_T$ ($p_0 = 0.15$, $L = 1$, $TOX = 1.8$) 122
C.3 Bias of $\hat{k}_T$ ($p_0 = 0.15$, $L = 1.5$, $TOX = 1$) 123
C.4 Bias of $\hat{k}_T$ ($p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$) 124
C.5 Bias of $\hat{k}_T$ ($p_0 = 0.05$, $L = 1$, $TOX = 1.8$) 125
C.6 Bias of $\hat{k}_T$ ($p_0 = 0.05$, $L = 1$, $TOX = 1.8$) 126
C.7 1-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1$, $TOX = 1$) . 128
C.8 1-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1$, $TOX = 1.8$) 129
C.9 1-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1.5$, $TOX = 1$) 130
C.10 1-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$) 131
C.11 1-sided Type I error and Power Comparisons ($p_0 = 0.05$, $L = 1$, $TOX = 1.8$) 132
C.12 2-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1$, $TOX = 1$) . 134
C.13 2-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1$, $TOX = 1.8$) 135
C.14 2-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1.5$, $TOX = 1$) 136
C.15 2-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$) 137
C.16 2-sided Type I error and Power Comparisons ($p_0 = 0.05$, $L = 1$, $TOX = 1.8$) 138
D.1 Bias of $\hat{k}_T$ under Ho ($p_0 = 0.15$, $L = 1.5$, $TOX = 1$) 140
D.2 Bias of $\hat{k}_T$ under Ho ($p_0 = 0.15$, $L = 1$, $TOX = 1.8$) 141

D.3 Bias of $\hat{k}_T$ under Ho ($p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$) 142
D.4 Bias of $\hat{k}_T$ under Ha ($p_0 = 0.15$, $L = 1.5$, $TOX = 1$) 143
D.5 Bias of $\hat{k}_T$ under Ha ($p_0 = 0.15$, $L = 1$, $TOX = 1.8$) 144
D.6 Bias of $\hat{k}_T$ under Ha ($p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$) 145

GLOSSARY

CARCINOGENICITY: the ability of an agent to cause cancer / cancerous lesions / tumors; alternately: tumorigenicity, oncogenicity

CA TEST: Cochran-Armitage trend test; forms the basis for the poly-k test

CDC: Centers for Disease Control

FDA: Food and Drug Administration

MH TEST: Mantel-Haenszel (alternately, Hoel-Walburg) test; adjusts for time by analyzing data by time intervals

MURINE: relating to rats or mice; e.g. a murine model for study of a biological effect

NECROPSY: an autopsy performed on an animal

NIEHS: National Institute of Environmental Health Sciences; subdivision of the NIH

NIH: National Institutes of Health

NTP: National Toxicology Program; a multi-agency U.S. program headquartered at the NIEHS (and also working with the FDA and CDC); evaluates substances of public health concern, develops methods, and communicates results and data to the public

TOXICOLOGY: scientific study concerned with identifying and quantifying the effects of poisonous substances

OCCULT: hidden from view; e.g. an occult tumor which is only discovered upon necropsy

POLY-K TEST: a generalization of the CA trend test, introduced by Bailer & Portier (1988) [3]

ACKNOWLEDGMENTS

Pursuing this PhD has been the largest undertaking of my life to date and its completion would not have been possible without the many forms of support I received along the way.

First and foremost, I would like to thank my advisor, Dr. Barbara McKnight, who shepherded me through several versions of this project, for your guidance, patience, and moral support throughout this journey. Your considerable knowledge and insights have been invaluable, and I have enjoyed our discussions, both academic and personal, during this time.

To my committee members, Dr. Brian Leroux and Dr. James Hughes, I owe much appreciation for your excellent feedback and advice, and for taking the time to meet regularly with me during the final stages of my work. Thank you also to Dr. Chris Li, for serving as committee GSR, and for donating your time.

Much belated thanks are due to Dr. David Schoenfeld, who encouraged me, back in my days at MGH Biostatistics, to pursue a PhD.

Thank you to the University of Washington for the opportunity to conduct this research. I have been very grateful for the funding I received to assist in completing this degree, including the generous NIDCR Ruth Kirschstein training grant.

Finally, thank you to my family and friends, who have been nothing but supportive during the years I spent in graduate school. Especial thanks to my husband, Jon, and daughter, Allie, who endured creative shifts in what it means to have work-life balance as I finalized this project. Thank you to my parents-in-law, whose help made many weekend work sessions possible. And much thanks to my parents, who have been there from the beginning, teaching me to prioritize education, and always believing I could achieve what I set out to accomplish.

DEDICATION

To Jon and Allie.

Chapter 1

INTRODUCTION

Each year, hundreds of new substances are released into the environment, ranging from food additives and personal care products to industrial chemicals, with unknown effects on public health [41]. It is therefore of interest to examine exposures that may be hazardous to humans. Animal carcinogenicity experiments are conducted by private entities and government agencies such as the National Toxicology Program (NTP) to investigate whether a substance causes cancer. Our research examines existing and newly proposed hypothesis testing methods for this setting. We also consider the impact of experimental design on tests' performance characteristics.

1.1 Carcinogenicity Experiments

The lifetime tumorigenicity experiment is one type of toxicology experiment in which murine (mouse and rat) carcinogenicity models are used to investigate whether a substance causes cancer at different doses over a 2 year life span. In its simplest form, such an experiment concludes with a single terminal sacrifice of the animals surviving to the end of the study [28]. The highest doses given in these experiments are often toxic or might affect tumor lethality, inducing differences in mortality across treatment groups. Since most tumors are occult (i.e. not detected until necropsy) and it is necessary to sacrifice the animal, or wait until natural death, in order to detect the presence of cancer [14], producing unbiased tests for differences in tumor incidence is challenging [24].

1.2 Existing Methods

Various methods have been proposed for analyzing such data to detect differences (or trends) in cancer across dose groups. These include Tarone's test (a log rank test for trend), which works well when there is a surrogate for time to tumor onset, such as with non-occult or instantly-lethaly tumors [40]; the Hoel-Walburg (aka Mantel-Haenszel) test, which uses prevalence estimates within time strata and is useful for non-lethal tumors; and the linear trend score test from logistic regression, which allows for continuous adjustment for time in the non-lethal tumor setting [24]. It is well-recognized that many statistical methods can perform poorly (as measured by type I error rate and power) in lifetime experiments when their assumptions are not met; this often occurs when there are differences in mortality induced by treatment toxicity or tumor lethality [3, 12, 13]. The Cochran-Armitage (CA) trend test (discussed below) is not affected by tumor lethality, but has poor performance in the presence of treatment toxicity. The poly-3 test, based on the CA test, was developed to address this problem.

1.2.1 Cochran-Armitage Trend Test

The Cochran-Armitage (CA) test is a regression model trend test which uses a regressor variable to account for ordering among treatment groups [2]. For groups $i = 0, ..., I$, let y_i be the number in group i with tumor present at death, z_i the dose level for group i, with $\bar{z}_w$ a weighted average of the dose levels, $p_i = y_i/n_i$ the proportion of group i with tumor, and $p = \sum_i y_i / \sum_i n_i$ the proportion of animals that developed a tumor during the experiment. The CA test statistic is then calculated as:

$$Z_{CA} = \frac{\sum_i n_i(p_i - p)(z_i - \bar{z}_w)}{\sqrt{p(1-p)\sum_i n_i(z_i - \bar{z}_w)^2}} = \frac{\sum_i y_i z_i - p\sum_i n_i z_i}{\sqrt{p(1-p)\left[\sum_i n_i z_i^2 - \frac{(\sum_i n_i z_i)^2}{\sum_i n_i}\right]}} \tag{1.1}$$

By assuming that, absent carcinogenicity, all animals are at equal risk of developing cancer, the CA test encounters problems when there are differences in intercurrent mortality among treatment groups. With (non-oncogenic) treatment toxicity, animals in higher dose groups are more prone to dying before cancer onset, which leads to conservative behavior and diminished power [3].

1.2.2 Poly-k Survival Adjusted Trend Test

The poly-3 (or more generally, the poly-k) test developed by Bailer & Portier [3] is a generalization of the CA trend test which modifies the number at risk according to animals' time on study. This adjusted number at risk is defined as:

$$n_i^* = \sum_{j=1}^{n_i} \alpha_{ij}, \text{ with animal-specific } \alpha_{ij} = \begin{cases} 1 & \text{if animal died with tumor} \\ (t_{ij}/t_{max})^k & \text{if not} \end{cases}$$

The new statistic takes the form of expression (1.1), with n_i^*-based values filling in for n_i, p_i, and p; $p_i^* = y_i/n_i^*$ and $p^* = \sum_i y_i / \sum_i n_i^*$ are the proportions of animals with tumor, adjusted for time-at-risk. This formulation assumes that the amount of time each tumor-free animal is at risk is a function of its time to death raised to the power k. Setting $k = 0$ reduces the test back to the CA.

Bailer & Portier calculated the null variance in the original poly-k test statistic simply as $\text{var}(p^*) = p^*(1-p^*)/n_i^*$. This assumes a fixed denominator to the event rate, ignoring the random component introduced by the computation of n^*, thus underestimating the variance and inflating the test statistic. Bieler & Williams (1993) derived a less biased variance estimator for this statistic using the delta method. (Details available in Appendix A.1.) The Bieler & Williams form of this statistic is shown in (1.2), and is now standard wherever the

test is applied [5].

$$Z_p = \frac{\sum_i \frac{n_i^*}{n_i} y_i z_i - \frac{\sum_i (n_i^*/n_i) y_i}{\sum_i (n_i^*)^2/n_i} \sum_i ((n_i^*)^2/n_i) z_i}{\sqrt{\frac{\sum_{ij}(r_{ij}-\bar{r}_i)^2}{n-(I+1)} \left[\sum_i \frac{(n_i^*)^2}{n_i} z_i^2 - \frac{(\sum_i ((n_i^*)^2/n_i) z_i)^2}{\sum_i (n_i^*)^2/n_i}\right]}}, \text{ where } r_{ij} = y_{ij} - p^* \alpha_{ij} \qquad (1.2)$$

If the shortcomings of the CA test are due in part to the fact that all animals are treated as spending equal amounts of time at risk (even when they shouldn't be), adjusting the number at risk based on additional information is a natural modification to make. For a tumor-bearing animal, death prior to 2 years is not a censoring event, so no adjustment is made. However, animals who died without tumor provide information up to their death time, not through t_{max}. Thus the poly-k test adjusts each non-tumor-bearing animal's contribution to the number at risk by $\alpha_{ij} = \frac{H(t_{ij})}{H(t_{max})}$, i.e. the ratio of the cumulative hazard rate up to each animal's actual time of death to the potential cumulative hazard that would have accrued had this animal survived to the end of study. This adjustment to the animals at risk will clearly depend on the choice of distributional assumptions, through the cumulative hazard. For example, for a constant hazard, $H(t) = \lambda t$, and so $\alpha_{ij} = t_{ij}/t_{max}$; for a linear hazard ($H(t) = \lambda t^2/2$), $\alpha_{ij} = (t_{ij}/t_{max})^2$; and of course these are simply reductions of the case of a Weibull hazard with shape parameter $k > 0$ (with $H(t) = \lambda^k t^k$), for which this quantity is precisely $(t_{ij}/t_{max})^k$, the α_{ij} chosen by Bailer & Portier. Citing earlier results from fitting Weibull models to historical control data [35], Bailer & Portier suggest setting $k = 3$ for this application [3].

1.3 Literature on the poly-k test

In the absence of treatment toxicity, the poly-3 test has been shown to maintain type I error rate even when the underlying Weibull hazard has a different shape parameter from the $k = 3$ assumed [3, 5, 11]. Among methods that keep size close to the nominal level, Kodell et al. report that the poly-3 achieves the best power [21]. Under many scenarios, Bailer &

Portier report that the CA trend test and poly-3 perform comparably well, and both are robust to changes in tumor lethality [3]. Robustness to tumor lethality makes sense, since both tests depend on detecting the presence of tumor, and tumor lethality only affects what happens after the tumor occurs.

Under differential intercurrent mortality, competing risks may not be independent of treatment group, with animals at higher dosages therefore more likely to die before tumor onset. Assuming the Weibull assumption is correct, the poly-3 at-risk adjustment is designed to correct for this bias; since time to tumor onset cannot be directly observed, α_{ij} relies on the correctness of the Weibull(k) assumption to model it. In the presence of treatment toxicity, Bailer & Portier showed that the poly-3 test outperforms the CA test for cancers with a moderately-low to high ($4.6 - 19.1\%$) baseline tumor rate; the underlying tumor hazards included Weibulls with $k \in (0.75, 2.7, 5.5)$. The CA test is conservative in these cases (with type I error at $1.2 - 1.9\%$) whereas poly-3 maintains closer to nominal type I ($3.8 - 4.2\%$ for moderate tumor rates and $5.0 - 5.6\%$ for high tumor rates). This is reassuring since it is precisely the sort of scenario that Bailer & Portier's modification to the CA test aims to adjust for. Under all simulation settings, the poly-3 has either higher or comparable power to the CA trend test.[3]

When the poly-k Weibull shape parameter assumption is incorrect, its adjustment to number of animals at-risk will either over- or under-estimate group-wise n_i^*. In the presence of treatment toxicity, this results in an asymmetry, with some treatment groups' at-risk numbers more biased than others. Bailer & Portier described the problem with toxicity (i.e. treatment lethality) as follows (note that their true underlying tumor hazard shape parameter is η_2): "If $\eta_2 < 3$...the factor $(t_{ij}/t_{max})^3$ will be smaller than it would be if η_2 were used as an exponent. Thus, the number at risk would be smaller using the exponent equal to 3, which in turn implies that the estimate of the probability of tumor onset would be larger than it would be if the true onset shape parameter were used." [3] The inflated test statistic

results in an inflated type I error rate from the poly-3 test. By analogous argument, if the true shape parameter is greater than 3, the poly-3 test will reject the null less frequently. Indeed, many authors have shown in simulation studies under treatment toxicity that when the true tumor hazard has $k < 3$, the poly-3 becomes anti-conservative, whereas using poly-3 when the true $k > 3$ yields conservative behavior; this effect becomes more pronounced for greater treatment toxicity [3, 5, 11, 18, 21, 26].

Bailer & Portier report that the CA test outperforms the poly-3 for high treatment toxicity when cancer is rare or late-onset; for rare cancers, the CA test is closer to nominal type I ($3.2 - 4.3\%$) than the anti-conservative poly-3 ($8.0 - 9.8\%$) in the same scenarios; the poly-3 becomes conservative when cancer is late-onset. This again is attributed to the effects of model misspecification: The authors note that the Weibull shape parameter used to generate the "rare cancer" data (1.2% baseline rate) was a great deal less than the $k = 3$ assumed by the poly-3 test (although they do not specify it exactly). To achieve late-onset cancer, they specify $k > 3$. [3] We note that the Bieler & Williams modification to the variance reduces type I error inflation for these rare cancer scenarios [5]. Using the Bieler-Williams version of the test statistic, Moon et al. (2003) show that, for a given misspecification of k, the failure of the poly-k to maintain size when intercurrent mortality varies according to treatment gets slightly worse with higher lifetime cumulative tumor probabilities (which they ranged from 0.05 to 0.30) [26].

Acknowledging that the choice of k parameter for the poly-k test is somewhat arbitrary without additional guidelines, and that model misspecification can lead to poor test performance when a treatment is toxic, Moon et al. (2003) [26] propose a method for estimating the value of k based on estimated lifetime cumulative tumor incidence rates, which they then apply in a "generalized poly-k test". In simulations of experiments applying 3 interim sacrifices, their new test exhibits type I error rates close to the poly-k test with correctly-specified k. Moon et al.'s method is discussed in greater detail in Chapter 3.

The existing literature that examines the performance of the poly-3 and competitor tests assumes Weibull tumor hazards in simulating tumor onset times (as well as times to death), simply varying the k parameter when assessing departures from the poly-3 assumptions. Bailer & Portier [3] model transition rate to tumor using Weibull, and death with a modified Weibull hazard that includes an offset term; Bieler & Williams [5], Kodell & Ahn [20], and Moon et al. [26] use the same scheme. Dinse [11], Peddada et al. [32], and Rahman & Lin [38] use Weibulls for both tumor and death onset hazard. Considering that Weibull tumor onset hazards are the parametric family assumed by the poly-k and by existing adaptive testing approaches (such as the one proposed by Moon et al. [26]), it is notable that previous authors have not violated this assumption in evaluating the performance of this test. Kodell (2012) notes this gap, saying that the predominance of the Weibull assumption in simulated data suggests "a need to evaluate both tests when data are simulated from distributions having different characteristics...the behavior of the poly-3 test under other distributional shapes is unknown" [18].

Given that the poly-3 test has become a standard for analyzing carcinogenicity experiments within agencies such as the National Toxicology Program (NTP) and the FDA [14, 18], and is routinely used by European pharmaceutical companies [18], further exploration of its limits and development of alternative tests is warranted.

1.4 Research Summary

Examining Bailer & Portier's at-risk adjustment α_{ij} under log-logistic and Gompertz hazards suggests that the consequences of some types of model misspecification may be worse than previously explored. In Chapter 2, we examine the performance characteristics (type I error and power) of the poly-k test and competitors when the underlying tumor onset hazard is non-Weibull. We compare log-logistic and Gompertz hazards to the usual Weibull settings. We also propose two new CA-based trend tests, which assume either log-logistic or Gompertz

hazards, and compare them with existing tests across different true underlying hazards.

Portier et al. (1986) declared from their analysis of historical control data from the NTP that $k = 3$ is justified in most experiments of this kind, and that $1 \leq k \leq 5$ is reasonable [35]. In practice, most applications of this test use $k = 3$; in particular, the current NTP statistical procedures guideline indicates $k = 3$ as the default for this test [30]. With some fore-knowledge of the likely shape parameter, an appropriate poly-k test can be chosen to improve performance. Moon et al. attempted this, but their generalized poly-k approach has several design features that might be improved upon. We developed an adaptive testing algorithm that estimates the poly-k parameter based on the data, using likelihood maximization in the Weibull family. In Chapter 3 we describe the new method in detail and examine how well it estimates k, comparing it with Moon et al.'s estimator. We also compare the power and type I error of our new test to existing tests, including the poly-k using Moon et al.'s k estimate, across a variety of Weibull and non-Weibull simulation settings.

In Chapter 4 we investigate the impact of serial sacrifice experimental design on the performance characteristics of the tests from Chapter 3.

Conclusions and future work are considered in Chapter 5.

All analyses and simulations were conducted using R versions 3.2.1, 3.2.3, and 3.3.0. [37]

Chapter 2

EFFECT OF VIOLATING THE WEIBULL ASSUMPTION ON THE PERFORMANCE OF THE POLY-3 TREND TEST

2.1 Introduction

The widely-used poly-k test, developed by Bailer & Portier (1988) [3] as a modification of the Cochran-Armitage (CA) trend test, was designed to avoid problems from differential mortality by adjusting for subjects' time-at-risk. This test makes the strong parametric assumption that underlying tumor hazards may be modeled as Weibull with shape parameter k (usually with $k = 3$). Specifically, the poly-k test statistic adjusts the number of animals as risk through the quantity $n_i^* = \sum_{j=1}^{n_i} \alpha_{ij}$. For tumor bearing animals, $\alpha_{ij} = 1$. For tumor-free animals, $\alpha_{ij} = \frac{H(t_{ij})}{H(t_{max})}$, where $H(t)$ is the cumulative underlying tumor hazard. The poly-3 test sets α_{ij} to $(t_{ij}/t_{max})^3$; it is readily seen that this value is obtained when the underlying hazard is a Weibull with shape parameter 3. The poly-k test was described in greater detail in section 1.2.2.

We noted in section 1.3 that previously published results examining the poly-k test and other methods for analysis of animal carcinogenicity data have relied on using Weibull hazards ($h(t) = \lambda^k k t^{k-1}$) to simulate tumor onset times. While the Weibull model is very flexible and easy to simulate, it is not clear that this is the only feasible choice to represent what is happening in nature. In the original paper proposing the poly-3, the authors admit that "It is unknown what could happen with other forms of the tumor incidence function" [3] and this gap has also been noted more recently by Kodell (2012) [18]. Poor behavior under non-Weibull data generation schemes would lead to concern about the suitability of the poly-k (and methods derived from it) for some types of data. Alternately, evidence that

data simulated under these alternative hazards does not in reality represent a very large departure from the assumptions (beyond what has already been considered under Weibull hazards) would indicate that, although there is room for improvement in these methods, it is not unreasonable to proceed with a Weibull family assumption.

Here we report our study of the behavior of the poly-3 "under less-than-ideal conditions", as suggested by Kodell [18]. Previous authors have seen that the poly-3 test behaves poorly in the presence of treatment toxicity when modeling assumptions are violated in the form of mis-specifying the shape parameter k of the underlying Weibull hazard. It was unknown whether the otherwise good behavior of this test relies on the true underlying hazard being, at the very least, in the Weibull family. The at-risk adjustment α_{ij} treats adjusted time at risk as a function of time to death simply raised to a power. The α_{ij}'s derived from a different assumed underlying hazard family can look very different; this is discussed in more detail in section 2.2.

Simulating data under hazards other than Weibull allows us to examine type I error and power for the poly-3 and competing methods in scenarios that may be scientifically reasonable, but that are outside the Weibull family of hazards examined so far. In selecting among non-Weibull hazards, we considered both the potential to add something new beyond typical Weibull shapes, as well as numerical tractability, and chose to examine the log-logistic ($h(t) = \lambda^k k t^{k-1}/[1 + (\lambda t)^k]$) and Gompertz ($h(t) = abe^{bt}$) hazards [22]. Figure 2.1 compares several Weibull, Gompertz, and log-logistic hazards.

Whereas the usual assumed range of Weibull shape parameters ($1 \leq k \leq 6$) in the literature includes hazards that are either monotone increasing or constant, the log-logistic allows for a hazard that rises higher earlier. Among human subjects, such a pattern might arise due to variability from genetics and other exposures; we see the initial rise in hazard as individuals more susceptible to cancer develop early tumors. Among inbred mouse and rat strains, genetic variability is low-to-nonexistent, which makes this less likely, though

competition for food in feeding experiments could induce such a hazard. By contrast, the Gompertz hazard allows us to model failure rates that increase exponentialy with time. Since the poly-k test treats time as being raised to a power k, we expected that the poly-k would perform poorly against Gompertz data that accumulates most events only at the end of the animals' natural lifespans.

In addition to considering different hazards for data simulation, we also consider how CA-based tests that rely on log-logistic or Gompertz distributional assumptions would compare against the poly-k under different types of true underlying hazards. Our new tests are described in section 2.2, details of the simulation study are outlined in section 2.3, and results are presented in section 2.5.1. We also describe the selection of a real-world example data set to which we apply these methods (sections 2.4 and 2.5.2).

2.2 Alternate Cochran-Armitage-Based Tests

We implement two new hypothesis testing methods using the same Cochran-Armitage basis as the poly-k, but with different parametric assumptions. Recall from section 1.2.2 that the adjustment factor to number at risk used by Bailer & Portier was $\alpha_{ij} = \frac{H(t_{ij})}{H(t_{max})} = (t_{ij}/t_{max})^k, k > 0$, obtained by using the Weibull cumulative hazard $H = \lambda^k t^k$. In principle, any choice of parametric cumulative hazard function $H(t)$ can be substituted into this quantity. In this way, we obtain two survival-adjusted tests, the $\text{CA}_{\text{log-logistic}}$ and $\text{CA}_{\text{Gompertz}}$, that respectively assume log-logistic and Gompertz hazards.

Like the poly-k test, the $\text{CA}_{\text{log-logistic}}$ and $\text{CA}_{\text{Gompertz}}$ tests are defined by a choice of parameters, with the log-logistic α_{ij}'s (and hence the corresponding new test) a function of 2 parameters ($\frac{H(t_{ij})}{H(t_{max})} = log(1 + (\lambda t)^k)/log(1 + (\lambda t)^k)$), and those derived under Gompertz a function of one parameter ($\frac{H(t_{ij})}{H(t_{max})} = (e^{bt_{ij}} - 1)/(e^{bt_{max}} - 1)$). We incorporate standard errors based on the delta method, as derived for Bieler & Williams' version of the poly-k [5]. (Further detail provided in Chapter 1 (section 1.2.2), with particulars on the delta method

approach to approximating variance in Appendix A.1.)

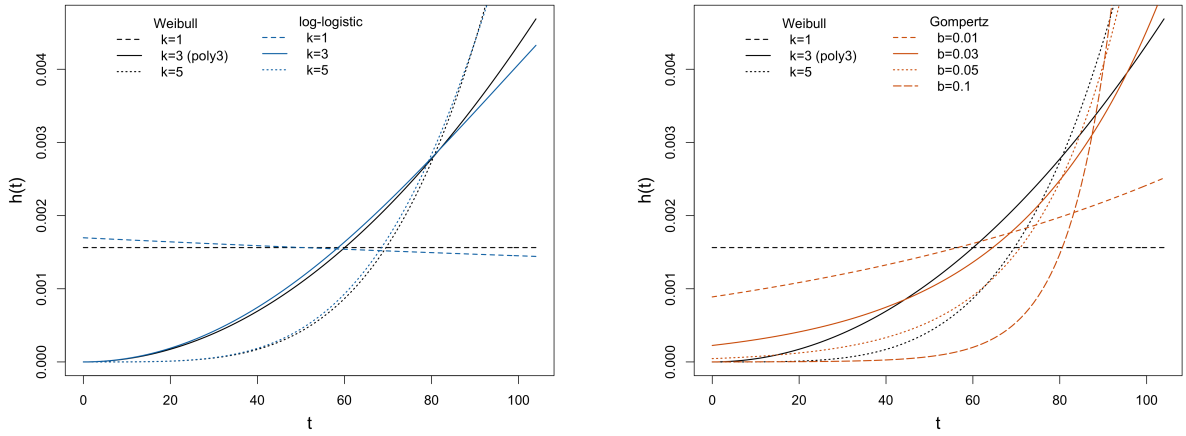


Figure 2.1: Comparison of Hazards

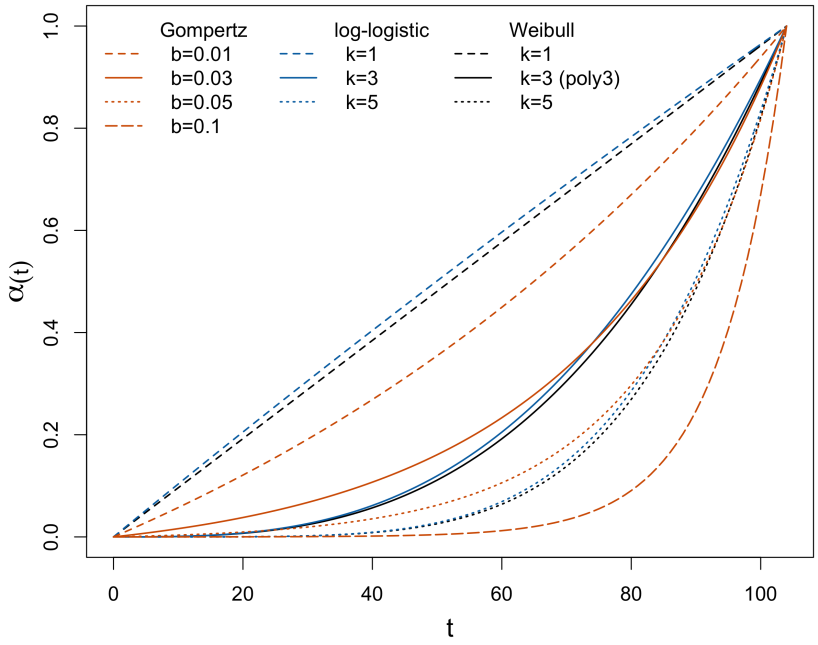


Figure 2.2: Animal-specific time-at-risk quantities α_{ij} for different hazard assumptions

In Figures 2.1 and 2.2, for each k we chose λ to yield a 0.15 probability of a tumor for animals surving to $t_{max} = 104$. For comparable k, the log-logistic and Weibull hazard curves look similar (Figure 2.1 left), but with key differences in curvature. For instance, for $k = 3$, log-logistic tumor hazard is higher than Weibull for much of the experiment, but is overtaken by the beginning of the last quarter. Comparing Weibull versus Gompertz (Figure 2.1 right) indicates potentially greater differences in the range of hazards. Gompertz with b=0.1 is included as an extreme example, which represents a vast majority of tumors occuring very near the end of the animal lifespan (or the duration of the experiment); this is, however, an unlikely scenario and corresponds roughly to a Weibull with $k = 100$.

In order to get a sense of how much such shape differences might translate to an effect on parametric tests' performance, it is helpful to compare the corresponding time-at-risk adjustments. Figure 2.2 shows the α_{ij} for a rat dying without tumor at a given time t_{ij} for a range of underlying hazard assumptions. If, for instance, the true tumor onset hazard is Gompertz(0.1) (orange dashed line at the bottom of Fig. 2.2), we see that the poly-3 test (solid black line) would overestimate the time-at-risk for any tumor-free rat dying before the end of study. When intercurrent mortality differs among treatment groups, such a bias in α_{ij} will incorrectly elevate the expected number of events in higher-dose groups relative to lower-dose groups, and the test would have poor power and behave conservatively. Similarly, if the true underlying hazard is indeed Weibull-3, the poly-3 would perform well, whereas the $\mathrm{CA}_{\mathrm{Gompertz}}$ with $b = 0.1$ would greatly underestimate animals' contribution to number-at-risk; in the presence of treatment toxicity, this will dampen the expected numbers of events in the higher-dose groups relative to the lower-dose, which can result in inflated type I error for the test. By comparison, to the left of the Weibull-3 line in Figure 2.2, we see that $\mathrm{CA}_{\mathrm{Gompertz}}$ with $b < 0.03$ will tend to overestimate animals' contribution to number-at-risk when the true data are Weibull-3; however, regardless of parameter choice, they will never be as biased as the poly-1 test (leftmost black dashed line). (It is easily shown using Taylor

expansion that as $b \to 0$, the limit for the Gompertz α_{ij} function is t_{ij}/t_{max}, i.e., the Weibull $k = 1$ line.) We can make similar comparisons among Weibull and log-logistic assumptions (Figure 2.2 blue lines).

These plots suggest that under some parameter choices for underlying log-logistic or Gompertz tumor hazards, we will see worse type I error or power from the poly-3 test than has been seen from simulating data under Weibulls with the wrong k-parameter. The newly proposed $\text{CA}_{\text{log-logistic}}$ and $\text{CA}_{\text{Gompertz}}$ tests are likely to perform better in those extreme scenarios, but are of course also prone to perform poorly when their own parametric assumptions are violated.

2.3 Simulation Study

Notation

- $i \in (1, ..., I)$ dose groups with $j \in (1, ..., n_i)$ animals per group; dose z_i
- t_{ij} time to death
- s_{ij} time to tumor
- $h^T(s)$ hazard of tumor onset
- $h^{DOC}(t)$ hazard for death from other (non-tumor) causes when no toxicity
- $h^{DFT|T}(t, s)$ conditional hazard for death from tumor
- $H(t) = \int_0^t h(u)\, du$ cumulative hazard
- $S(t) = \mathrm{e}^{-H(t)}$ survival function

We simulate a 2-year murine terminal sacrifice experiment with 4 dose groups (0 (control), 0.25, 0.5, and 1) and 50 animals per treatment arm. Each scenario is run with 10,000 replications. Times to tumor are modeled using either Weibull, log-logistic, or Gompertz hazards. The range of Weibull shape parameters ($0.5 \leq k \leq 10$) was selected based on the previous literature (common range $1 \leq k \leq 5$ [35]) plus more extreme low and high values for comparison; a comparable range of parameters was used for log-logistic ($0.5 \leq k \leq 10$) and Gompertz ($0.005 \leq b \leq 0.1$) simulations, representing a range of departures from the usual Weibull assumption. Once a shape parameter is selected, the second parameter is constrained by our choice of cumulative tumor incidence by 104 weeks in the control group ($p_0 = S^T(s = 104|k, \lambda)$). We consider cancers with p_0 of about 0.05 (rare), 0.15 (somewhat common), and 0.3 (common). Under the null hypothesis the cumulative tumor probability for the highest dose group is equal to the control group ($H_0 : p_1 = p_0$). There are two options for $H_a : p_1 = 2p_0$ or $p_1 = 3p_0$. Detailed settings are available in Table B.1 in Appendix B.1.

Times to death are modeled using Weibull hazards. The competing risk death hazard (h^{DOC}) is a Weibull($k = 2, \lambda = 0.007$), with a study-end competing risks survival rate ($S^{DOC}(t = 104)$) of 0.6 [26].

We vary tumor lethality (extra risk of death due to presence of tumor) and treatment toxicity (a dose-associated mortality modifier), which are known to affect many tests' performances [11]. Our lethality is defined as the ratio $\mathscr{L}(t, s) = (h^{DFT|T=s}(t, s)+h^{DOC}(t))/h^{DOC}(t)$, which represents the greater risk of death with tumor (given that a subject has developed a tumor at time s) over the competing risks death hazard (death of other causes). If we allow $h^{DFT|T}$ to depend on t and s only through h^{DOC}, we can rewrite this expression in terms of a constant lethality parameter L_{tum}: $h^{DFT|T=s}(t, s) = (L_{\text{tum}} - 1)h^{DOC}(t - s)$. By using this formulation, we have made lethality depend on the shape of the hazard function. Intuitively, $L_{\text{tum}} = 1$ (no tumor lethality) when $h^{DFT|T}(t, s) = 0$, i.e. there is no extra hazard of death due to tumor occurrence. Data were generated with $L_{\text{tum}} \in \{1, 2, 3, 5, 10\}$.

To simulate times to event, a uniform random variable is generated and set equal to the CDF associated with the hazard for that event. Time is then easily back-calculated. Each animal has a randomly generated time to tumor (T^T) and time to death from non-tumor causes (T^{DOC}); the hazards used for death or tumor onset were described above. If $T^{DOC} > T^T$, then the animal is classified as having developed a tumor. For animals who develop a tumor, we use $h^{DFT|T=s}(t, s)$ (hazard as defined above) to obtain $T^{DFT|T-T} = T^{DFT|T} - T^T$, the competing risk time to death from tumor since tumor onset. To this we add the previously obtained T^T, resulting in the time to death from tumor since study start: $T^{DFT|T} = T^T + T^{DFT|T-T}$.

Toxicity is modeled as a dose-associated modifier, which allows us to generate dose group-specific death times ($T^{DOC}_{z_i}$) from the distribution with hazard $h^{DOC}_{z_i}(t) = L^{z_i}_{tx} h^{DOC}_{z_0}(t)$. Since doses range from 0 (control) to 1 (highest dose), L_{tx} is the hazard ratio of death from other causes, comparing the highest dose to the control group. Note that this dose-related toxicity does not factor into our above-described tumor lethality relationship, and as such we presently assume no interaction between dose and tumor lethality. In our simulations, data was generated with $L_{tx} \in \{1, 1.5, 2, 3, 5\}$.

The dose effect on tumorigenicity is applied as a scaling parameter on the control group tumor hazard. If the control group hazard of tumor is described by the function $h^T_{z_0}(t)$, then our tumor hazard for treatment groups $i = 1, ..., I$, with associated doses z_i, is defined as $h^T_{z_i}(t) = (1 + \eta z_i) h^T_{z_0}(t)$, with dose-effect parameter η. We again let p_0 be the cumulative tumor probability through t_{max} for the lowest dose (control) group, and p_1 be this quantity for the highest dose group (assuming no toxicity effects or sacrifices). It is easily shown that, for any choice of underlying hazard, η is simply a function of these cumulative tumor probabilities: $\eta = log(1 - p_1)/log(1 - p_0) - 1$. When $p_1 = p_0$, as under the null hypothesis, $\eta = 0$ and it follows that $h^T_{z_i}(t) = h^T_{z_0}(t)\ \forall i$.

Data are generated under the null hypothesis ($\eta = 0$) to compare type I error and under several alternatives ($\eta > 0$) to compare power. Since some tests do not achieve nominal level in some settings, for power comparisons, tests are calibrated to have 0.05 size (out to 4 significant figures). We examined these performance characteristics for the 1- and 2-sided versions of our newly proposed tests, the $\mathrm{CA}_{\text{log-logistic}}$ and $\mathrm{CA}_{\text{Gompertz}}$, comparing with common hypothesis tests for this setting, including the poly-k with Bieler-Williams variance [3, 5], Cochran Armitage [2], log rank test for trend [40], Hoel-Walburg (Mantel-Haenszel) with 4 intervals, and logistic score (adjusting for continuous time with linear and quadratic terms) [12] tests.

2.4 Data Set Example

To illustrate the application of these tests, we sought real-world carcinogenicity data from the National Toxicology Program (NTP). We obtained an archive containing 3082 studies, of which 574 were in the form that we are examining: experiments of approximately 2 year duration (714-742 days) with 3 or 4 dose groups and 50 or 60 animals per treatment group [29]. Of interest would be an experiment with high differential intercurrent mortality (treatment toxicity), in which we would also expect reasonable power for detecting a carcinogenic effect. The NTP archive covers a wide range of neoplastic and non-neoplastic lesions; our focus for the purpose of this illustration was liver neoplasms, which are commonly observed in murine models.

To examine toxicity, we identified 392 studies with a large sample of non-tumor-bearing animals that died of natural causes; this was defined as having at least half the overall sample and half the highest dose group tumor-free, with at least 15% of tumor-free animals dying of natural causes (overall and highest dose). For each study, a Cox proportional hazards model was fit to data on the subset of tumor-free animals who died of natural causes, to model time to death over treatment group. Given the large number of experiments, a great variety

of dosing schemes was present, making it difficult to compare effect sizes among studies; to put everything on a common scale, treatment dose was redefined in reference to the lowest non-control dose, by dividing all doses by the smallest non-0 dose. An estimated hazard ratio of $2 < HR < 4$ ($p < 0.005$ for H_0:HR=0) was defined as suggesting strong toxicity but avoiding the highest toxicity levels that are more likely to result in a failed carcinogenicity experiment. Just 35 studies met this toxicity criterion.

Among this final set of candidates, we wish to select a study that is reasonably powered to examine tumorigenicity. Referring to simulated uncalibrated power for the poly-k, $\text{CA}_{\text{log-logistic}}$, and $\text{CA}_{\text{Gompertz}}$ tests, it appears that, for any fixed underlying tumor hazard shape, toxicity level, or effect size, CA-based tests have better power when the cumulative tumor rate by end of study is $p_0 = 0.15$ (versus rarer cancers of $p_0 = 0.05$ or more common at $p_0 = 0.30$). To achieve a power close to 0.8 at high toxicity, it would be necessary for the highest dose group to at least triple the number of tumors by end of study compared to the lowest dose group. However, a doubling is more likely in a typical study. (See Figure B.1 in Appendix B.2 for plots of these power comparisons.)

We therefore seek experiments that look similar to a study with control group baseline cumulative tumor onset rate of 0.15, and a much higher highest-dose group tumor onset rate of 0.3 (doubling of cumulative tumors). It is reasonable to look for a study with 10-15% of control group animals having liver tumor, but since this set of studies was selected for evidence of toxicity, intercurrent mortality in the highest dose group was necessarily higher than in the control group. Simulated numbers of animals that would be alive and dead with tumor during such a high toxicity experiment suggest that for a true study-end cumulative tumor onset of close to 30%, we would in reality expect to see 17-20% of high-dose animals with tumors. There is 1 study among the 35 candidates that comes rather close to this power criterion, with 10% of control group animals and 18% of the highest-dose group observed with tumor. From our toxicity examination, the HR for death among tumor-free animals in

this experiment was 2.18.

The chosen data set is from a 2-year experiment of Pulegone on female rats, with 4 groups of 50 animals each. Pulegone is a flavoring and fragrance added to food, drink, and tobacco products, derived from the mint family, and is found in high concentration in pennyroyal oil. Pulegone was administered in corn oil by gavage 5 days/week, with doses by weight (doses $\in (0, 37.5, 75, 150)$ mg/kg). There were 106 terminal sacrifices, 92 natural deaths or moribund sacrifices, and 2 "dosing accident" deaths, which we classify as sacrifices. The highest-dose group was switched to vehicle placebo after the first 60 weeks of the experiment due to high morbidity and mortality; even so, none of the highest-dose group survived to the end of the experiment. [31]

We analyzed this data set for liver cancer and bladder cancer outcomes, and conducted hypothesis tests using the poly-k, $\text{CA}_{\text{log-logistic}} - k$, and $\text{CA}_{\text{Gompertz}} - k$ with $k \in (1, 3, 5)$.

2.5 Results

2.5.1 Simulation Study Results

We examined several combinations of 2-year cumulative background tumor onset probabilities for control (p_0) versus highest-dose (p_1) group; the rest of the dose groups' tumor onset hazards are determined by these endpoints. Modifying the cumulative tumor onset probability or the effect size η did not alter the relative performance of tests; patterns of tests' type I error and Power versus toxicity simply differed in scale. Therefore, we present only results from the simulations run with $p_0 = 0.15$ and (under H_a) $p_1 = 0.30$. Detailed simulation settings are provided in Appendix B.1 Table B.1.

Figures 2.3 and 2.5 show type I error and Power results for a subset of simulation settings, focusing on the CA-based tests: the poly-k, $\text{CA}_{\text{log-logistic}} - k$, and $\text{CA}_{\text{Gompertz}} - k/100$ tests.

These tests are not affected by tumor lethality, so we omit the (redundant) lethal tumor results here. For brevity and relevance, we show just the 1-sided test results. Each figure

contains a matrix of plots with columns defined by choice of true underlying tumor hazard (Weibull, log-logistic, or Gompertz), and rows defined by choice of true shape parameter within that family. Toxicity (L_{tx}) is represented on the horizontal axis of each sub-plot. The right-most column in each figure displays the hazard functions in each row for reference; tumor hazards within-row have analagous parameter settings and are arguably comparable.

Absent toxicity, all CA-based tests have very similar type I error and power (calibrated to type I error): it is evident in Figures 2.3 and 2.5 that the different tests overlap at the first point in each plot (toxicity=1). Regardless of underlying tumor hazard type or parameter setting, the CA-based tests reject below nominal type I, with size of about 0.04. Power varies over underlying tumor hazard, ranging from 0.46 to 0.55 in these examples. For 2-sided tests, power is diminished compared with 1-sided tests when there is no toxicity (see Appendix B.4 Figure B.5).

When the model is specified correctly, the poly-3 test maintains similar type I error across mild-to-moderate levels of toxicity (Figure 2.3 Weibull $k = 3$), but becomes slightly more conservative for high toxicity ($L_{tx} \geq 3$); its size ranges from 0.039 (at no toxicity) down to 0.031 (at highest toxicity). High toxicity results in even more conservative behavior (size of 0.025) when tumor incidence is higher ($p_0 = 0.30$, Appendix B.3 Figure B.2). This is a behavior noted by Bailer & Portier under some simulation settings [3]. Other CA-based tests exhibit the same behavior, e.g. $\text{CA}_{\text{log-logistic}} - 1$ (dashed blue line in Fig 2.3 row 1 column 2) and $\text{CA}_{\text{Gompertz}} - 0.03$ (dashed orange line in Fig 2.3 row 2 column 3).

Under model misspecification, all CA-based tests' type I error was more clearly not maintained under toxicity. When the true underlying hazard had a shape parameter less than that assumed by the test, type I error was inflated with increased toxicity (e.g. any of the solid or dotted lines in the first row of Fig 2.3). For true underlying hazards with larger shape parameters than assumed, tests became conservative with increased toxicity (see for example the poly-3, $\text{CA}_{\text{log-logistic}} - 3$, and $\text{CA}_{\text{Gompertz}} - 0.03$ in row 3 of Fig 2.3). Our

alternative CA-based tests have similar performance characteristics to the poly-k, performing well when their own modeling assumptions hold, and with analogous inflation of type I error for comparable departures from assumptions. Although the type I error of all "comparable" CA-based tests (i.e. poly-k, $\text{CA}_{\text{log-logistic}} - k$, and $\text{CA}_{\text{Gompertz}} - k/100$ for any given k) tended to behave similarly across toxicity, the Gompertz based test (and to a lesser degree the log-logistic based test) tended to be a little more conservative than the poly-k, especially at the highest toxicity (Figure 2.3). (Examining 2-sided test performance, we note that type I error also increases with toxicity, though not as much in most cases as for 1-sided tests, since the rejection region is split among two alternative hypotheses (Appendix B.4 Figure B.4).)

As for type of model misspecification and type I error, we see by comparing the first to the second column in Figure 2.3 that simulating data under log-logistic hazards did not produce significantly different behavior from that under comparable Weibull hazards. By contrast, the effect of toxicity is a bit attenuated under Gompertz hazards (column 3), particularly for the smallest shape parameter (first row $b = 0.01$). For underlying tumor hazards that are Gompertz ≥ 0.03, the effect of toxicity on type I error inflation is slightly greater.

The (size-adjusted) power of CA-based tests decreased with toxicity under almost all underlying tumor hazards (Figure 2.5). The effect of toxicity on tests' power becomes greater when the true tumor hazard is defined by a larger shape parameter: Using the poly-3 test as an example, under Weibull-3 or log-logistic-3 tumor hazards, power falls below 0.40 when toxicity$= 3$; under Weibull-5 or log-logistic-5 tumor hazards, power< 0.4 is observed at a lower toxicity of 2, and at toxicity$= 3$ power is < 0.35; whereas under Weibull-1, toxicity does not reduce poly-3 power very much. Model misspecification does not appear to be primary in this decline of power with toxicity. For any choice of underlying tumor hazard and toxicity level, all CA-based tests had similar power when calibrated for differences in level: tests defined by a smaller shape parameter performed slightly worse regardless of whether their assumptions are correct (e.g. Fig 2.5 row 1, column 2, dashed blue line). (Power for the

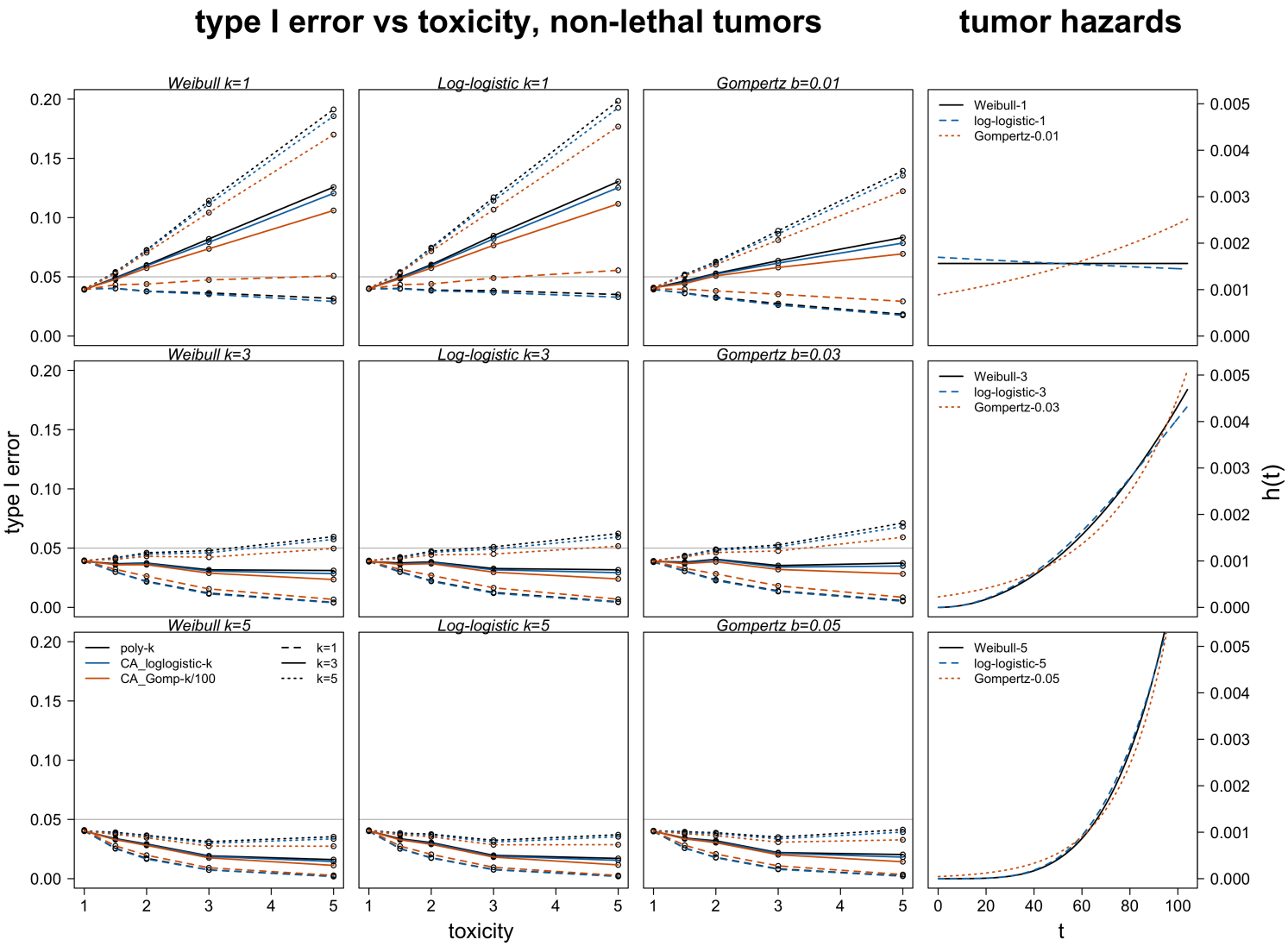


Figure 2.3: Data simulated under H_0 (type I error) with $p_0 = 0.15$, for nominal 0.05 level 1-sided tests

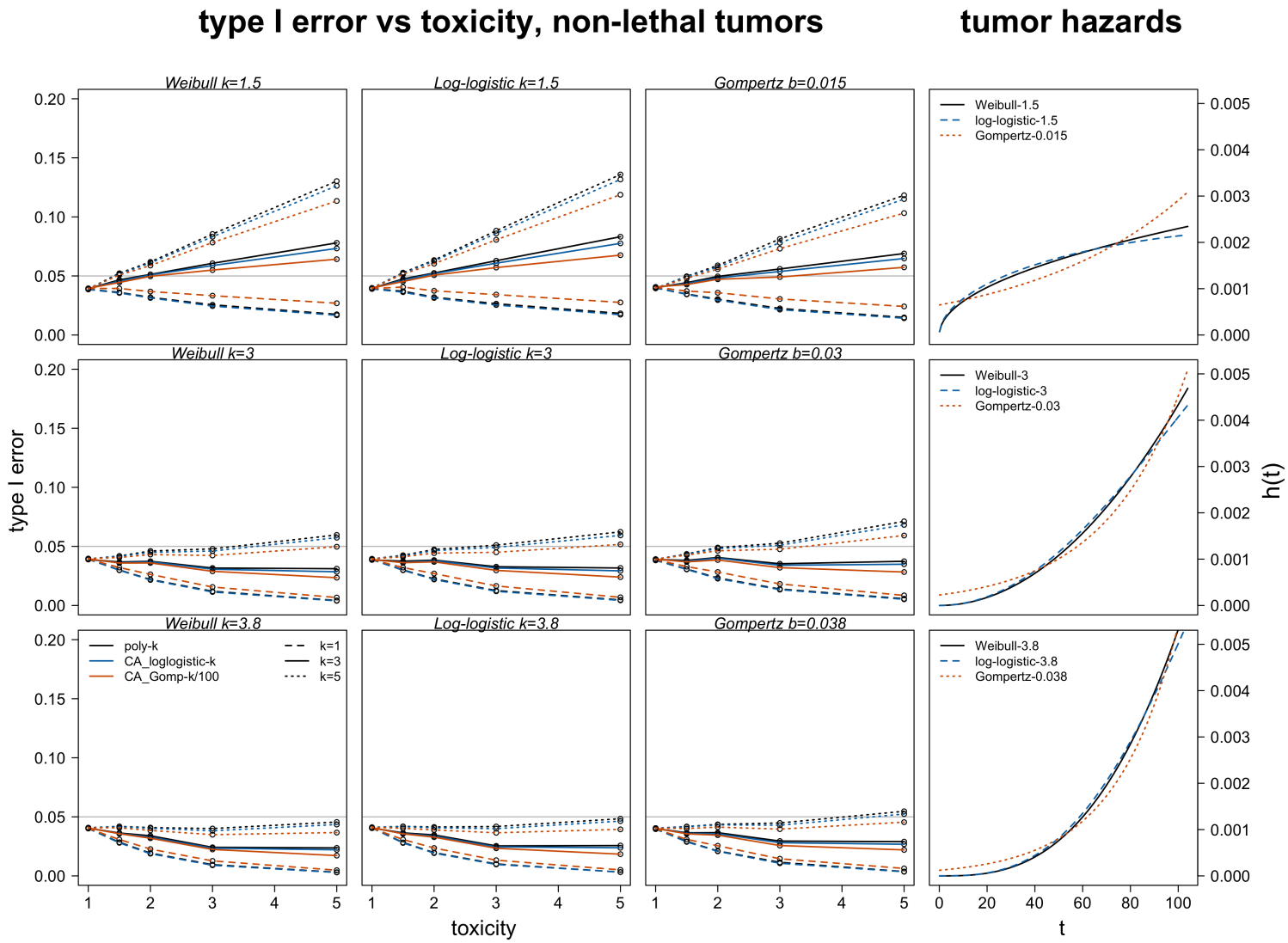


Figure 2.4: Data simulated under H_0 (type I error) with $p_0 = 0.15$, for nominal 0.05 level 1-sided tests (less extreme departures from poly-3 assumptions)

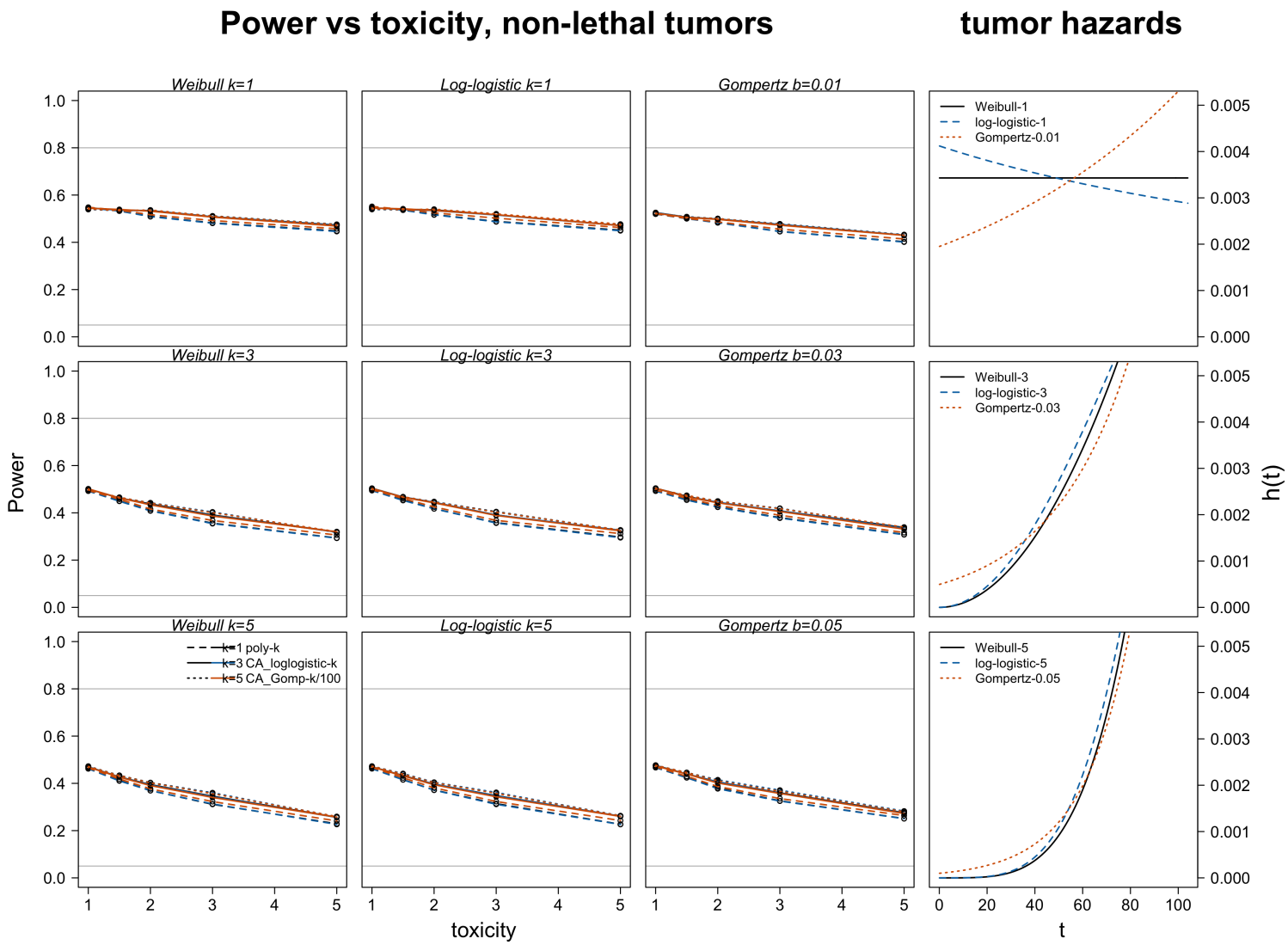


Figure 2.5: Data simulated under H_a (Power) with $p_0 = 0.15$, $p_1 = 0.3$, for 1-sided calibrated 0.05 tests

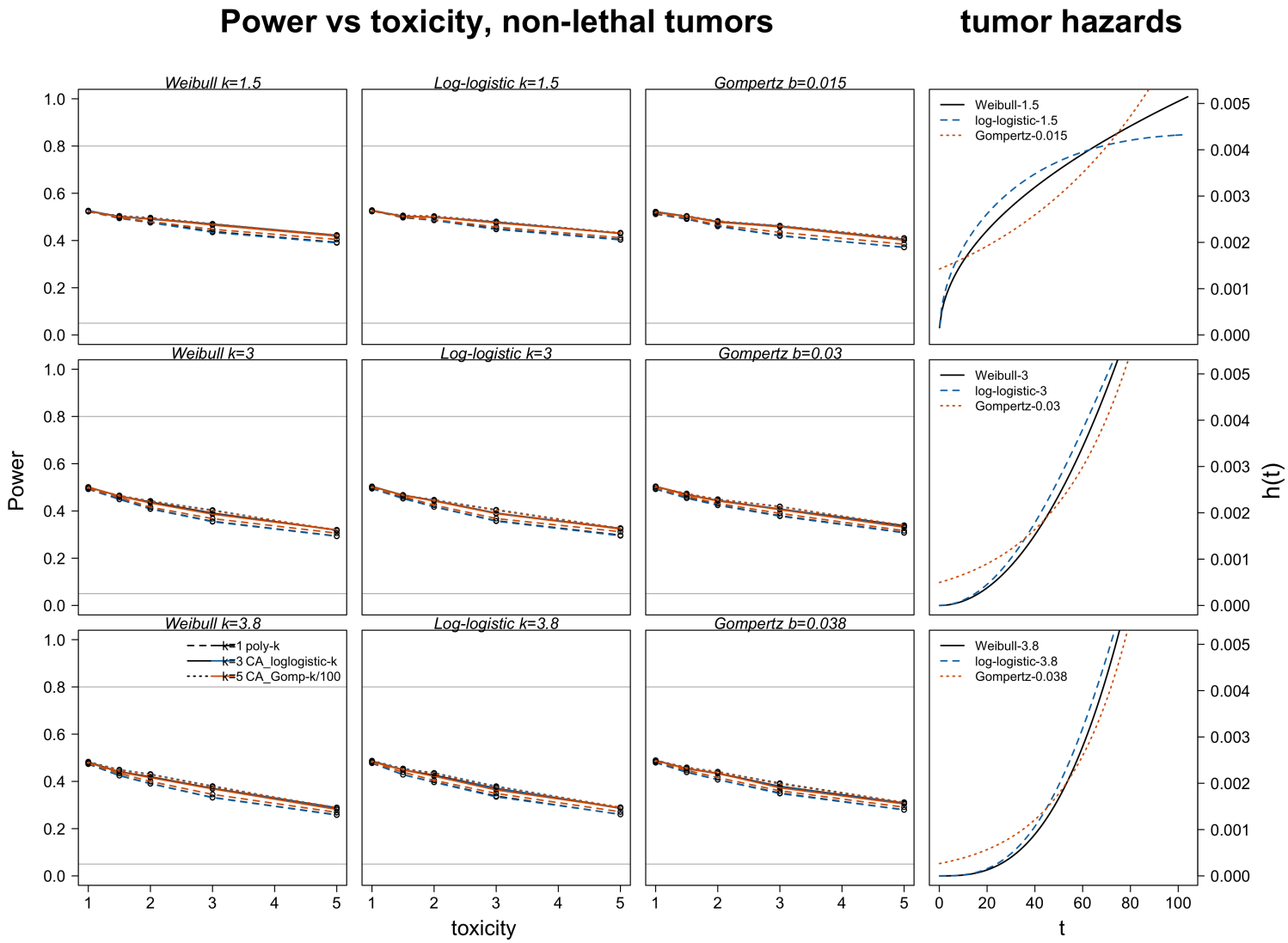


Figure 2.6: Data simulated under H_a (Power) with $p_0 = 0.15$, $p_1 = 0.3$, for 1-sided calibrated 0.05 tests (less extreme departures from poly-3 assumptions)

2-sided tests increased with greater toxicity under some conditions, likely due to rejecting the "wrong" side alternative (Appendix B.4 Figure B.5).)

How large model misspecification must be to affect poly-3 test level depends on the tumor incidence rate. (We ignore power as we have seen that model misspecification does not seem to influence the effect of toxicity on calibrated power.) At a low baseline cumulative tumor onset rate ($p_0 = 0.05$, see Appendix B.3 Figure B.3), simulating tumor onset hazards with a Weibull-1 or log-logistic-1 is required for the poly-3 to have inflated type I error at moderate levels of toxicity, where the type I error rate is > 0.06 at toxicity$= 3$. When the tumor onset hazard shape parameter gets slightly closer to the assumed 3, as with a Weibull-1.5, type I error only exceeds 0.06 for very high toxicity (type I error $= 0.055$ when toxicity$= 3$ and 0.0615 when toxicity$= 5$). For more common cancers ($p_0 = 0.15$), type I error inflation is greater for the same toxicity levels: When the true underlying hazard is Weibull-1, type I error is 0.06 at toxicity$= 2$ and up to 0.082 at toxicity$= 3$ (Fig 2.3); even under a lesser departure from the assumed shape parameter (Weibull-1.5), type I error is 0.061 at toxicity$= 3$ (Fig 2.4). For more common cancers ($p_0 = 0.30$, not shown), the toxicity thresholds are still lower for achieving the same type I error inflation.

We just noted that the change of type I error with treatment toxicity for given true shape parameter k takes similar form across different baseline cumulative tumor onset probabilities, simply varying in scale. Power, unsurprisingly, tended to be higher for larger simulated η, and was slightly more robust to the effect of toxicity than it was under smaller simulated η.

Unlike the CA-based tests, other competitor tests are affected by tumor lethality. To compare how lethality affects different tests, we simulated data under Weibull-3 (Figure 2.7). When there is no toxicity, we see that the logrank, Mantel Haenszel, score test, and even the original CA test (i.e. the poly-0), reject closer to the nominal 0.05 level than the CA-based tests; however, the MH and score tests show a small reduction in power when tumors are lethal. Among tests of the same size, the poly-3 test has the best power in the

absence of toxicity. The logrank test has higher power when tumors are lethal and there is no treatment toxicity, which makes sense since the test is expected to work well when we have a surrogate for tumor onset time, as with instantly lethal tumors [23].

In the presence of toxicity, the level and power of all of the CA-based tests (including the original CA) continue to be robust to the effect of lethality. Toxicity reduces the (original) CA test's power, which is expected since this was a main motivation for creating the poly-k test [3]. The log-rank test's performance is also affected by toxicity, with very inflated type I error for even low levels of toxicity, especially when lethality is low. The type I error rate of the score and MH tests is not changed by treatment toxicity, whereas their power diminishes with increased toxicity. This loss of power for the score and MH tests is exacerbated for higher levels of lethality. These findings are consistent with results published earlier in the literature comparing the poly-3 with competitor tests. [3, 21] These comparisons newly indicate that being based on the CA test, our new tests enjoy the same robustness to lethality as the poly-k does.

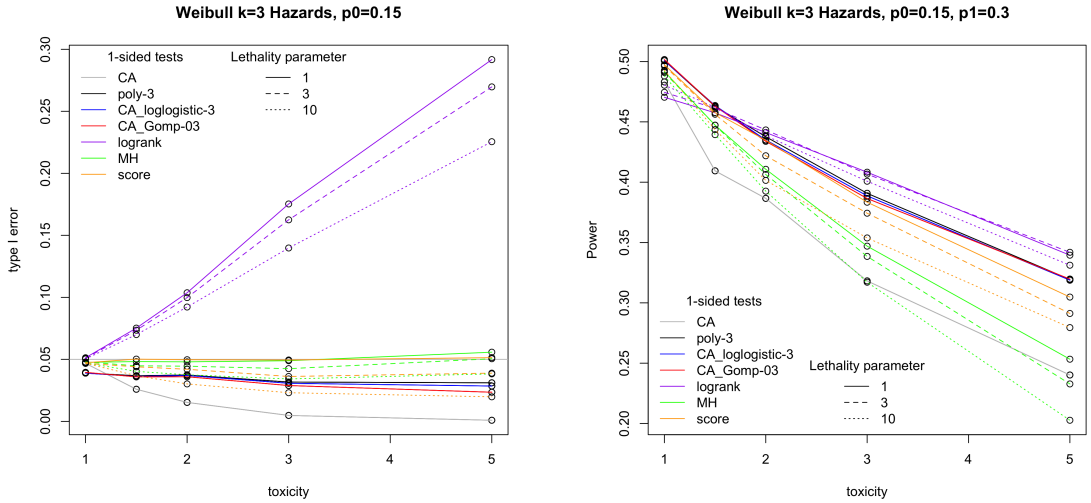


Figure 2.7: Comparison of CA-based tests with other competitors, by lethality

2.5.2 Data Set Example Results

Figure 2.8 (left) shows the cumulative incidence curves for death with liver cancer, and suggests a treatment effect on death with tumor. Particularly considering the toxicity of this particular agent to rats, this curve in part demonstrates the differential mortality among the treatment groups. For instance, the 150 mg/kg line goes up to probability 1 at the end because, by the time the last tumor was observed, all other animals in that treatment group had already died. Liver cancer incidence was 5/50, 5/50, 11/50, and 9/50 over the 4 dose groups, a noticeable increase in tumors for the higher dose groups compared with the lower doses. If we judge by the usual poly-3 test, this result is statistically significant (1-sided $p = 0.0125$).

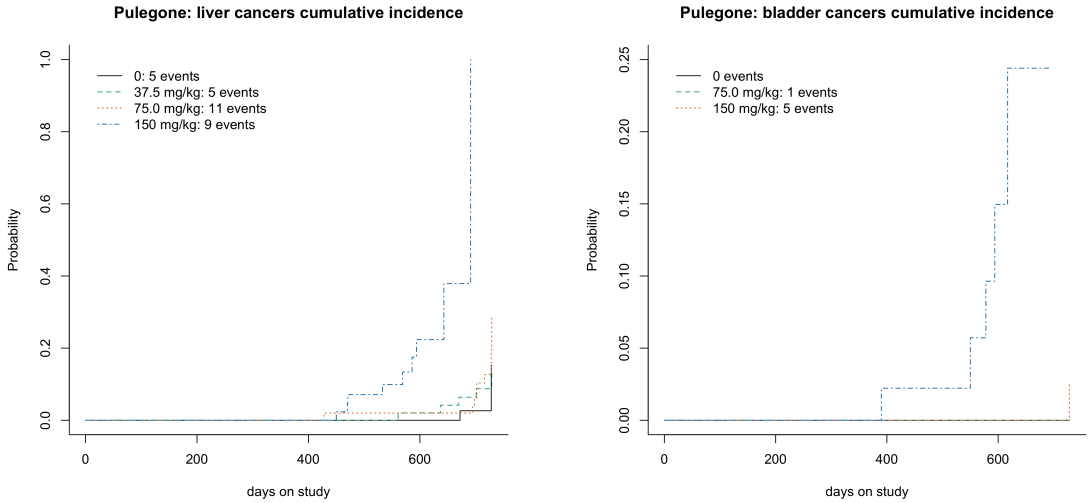


Figure 2.8: Data Set Example: Pulegone Cumulative Incidence Curves

In Table 2.1, we compare the poly-k and our competitor statistical tests. Under different modeling assumptions, statistical significance of course varies; most of these tests would lead us to reject H_0 at the 0.05 level, except for the test conducted under the assumption of log-logistic(1) tumor hazard. Interestingly, while we looked specifically at hepatocellular carcinoma, the NTP report on Pulegone appears to have grouped liver lesions together for

Table 2.1: Liver Cancer vs. Treatment Group

p-values by test	$k = 1$	$k = 3$	$k = 5$
poly-k	0.037	0.0125	0.0083
$\text{CA}_{\text{log-logistic}}$-$k$	0.0518	0.0475	0.049
$\text{CA}_{\text{Gompertz}}$-$k/100$	0.0079	0.0137	0.0148

Table 2.2: Bladder Cancer vs. Treatment Group

p-values by test	$k = 1$	$k = 3$	$k = 5$
poly-k	0.0019	0.0029	0.0059
$\text{CA}_{\text{log-logistic}}$-$k$	0.0019	0.0019	0.0019
$\text{CA}_{\text{Gompertz}}$-$k/100$	0.0098	0.0614	0.0741

this experiment: "Hepatocyte cellular alteration was the term used to describe a constellation of lesions in the liver." [31] Nevertheless, in agreement with our analysis, the NTP report did conclude, using the poly-3 test, that there was a significant increase in hepatocyte cellular alteration in the 75 and 150 mg/kg dose groups compared with the lower doses.

The NTP report on Pulegone also cited bladder cancer in particular, so we also examine this endpoint. Figure 2.8 (right) shows the cumulative incidence curves for death with bladder cancer. There are fewer bladder than liver cancers overall, but there is a much sharper trend with dose. In agreement with the NTP report, we find the same numbers of bladder tumors (papilloma or carcinoma) across the dose groups: 0, 0, 1, and 5. Table 2.2 compares the results of our statistical tests with the poly-k at different parameter values. With the exception of the Gompertz-based test, which deviates most from the usual assumed underlying hazard, our tests reject H_0 at $p < 0.01$. Using the poly-3 test, the NTP also concluded that there was "clear evidence of carcinogenic activity of pulegone in female F344/N rats based on increased incidences of urinary bladder neoplasms." [31]

2.6 Discussion

Our simulations suggest that our alternative CA-based modified tests can have better type I error rates than the poly-3 test under scenarios that are consistent with their own assumptions, but the improved performance is similar to what can be achieved by choosing a better k for the poly-k test itself. When the underlying tumor hazards are truly Weibull-3, these new modified tests unsurprisingly offer no advantage over the poly-3 in terms of maintaining size or power. These tests are sensitive to model misspecification in the presence of toxicity, just as the poly-k is.

In the presence of toxicity, when a CA-based test assumes a tumor onset hazard with shape parameter that is greater than the truth, i.e. when the test assumes a hazard shape with more late onset tumors than the truth, the result is inflated type I error. In this case, the test underestimates the α_{ij}'s (Figure 2.2), and so the effective number at risk (and expected number of events) in each group is smaller than it should be. When there is also differential intercurrent mortality due to toxicity, higher dose becomes inversely associated with time to death, and so the erroneous reduction in α_{ij}'s (expected events) is worse with higher dose. For a 1-sided test, this imbalance results in rejecting H_0 more frequently than 5% of the time. In the reverse scenario (in which the test assumes a tumor onset hazard with shape parameter that is less than the truth), the test will overestimate the α_{ij}'s, and to a greater degree for higher doses when toxicity culls higher dose animals, and so type I error is deflated. (For our 2-sided tests in this scenario (Appendix B.4 Figure B.4), we saw higher type I error with increased toxicity, as the test incorrectly detects an "effect" in the wrong direction.)

We found that CA-based tests with underlying hazards that were very similar nevertheless had slightly different type I errors, especially at high toxicity. For example, in the first plot in Figure 2.3, the poly-3 test has slightly higher type I error than $\text{CA}_{\text{log-logistic}} - 3$, which is slightly less conservative than $\text{CA}_{\text{Gompertz}} - 0.03$. This is likely due to differences in model

misspecification: even CA-based tests that are most similar to a given poly-k test are not making assumptions that are identical to the Weibull-k, and therefore their type I errors are not subject to the effect of toxicity in exactly the same way. For example, in contrast with the more gradual accrual of tumor onsets assumed under log-logistic or Weibull hazards, the $\mathrm{CA}_{\mathrm{Gompertz}}$ (with $b < 0.03$) assumes most events will occur later in the experiment, so its type I error is not as influenced by a toxicity-induced decline in high-dose tumor-bearing animals; the test effectively assumes that animals dying earlier in the study will tend not to be tumor-bearing. For the same reason, when the true underlying hazard is actually Gompertz, the type I error of all tests is less elevated with toxicity, since most animals killed early in the experiment have not yet had a chance to develop tumors, regardless of treatment group.

Since we calibrate for type I error, the effect of toxicity on power did not seem to be mediated by model misspecification so much as by issues of diminished sample size. Higher toxicity will lessen animals' contributed time at risk due to die-off, hence reducing power, no matter what the distributional assumptions. We saw this decline in power for all competitor tests (Fig. 2.7 right) as well as for the CA-based tests.

We made an effort to find just how bad the departure from modeling assumptions had to be for there to be a notable decline in the performance of the poly-3 test. For type I error inflation to be of concern for a somewhat common cancer, the minimum required departure from the Weibull-3 assumption was a Weibull-1.5 or log-logistic-1.5. For rarer cancers, this departure is closer to a Weibull-1. While this is not terribly close to $k = 3$, it is nevertheless within the Weibull($1 \leq k \leq 5$) range suggested by Bailer & Portier as being reasonable for these experiments [35]. The inflated error would be greater, of course, if the poly-3 test were not acting conservatively to begin with, as would happen with a larger sample size.

Whether or not an underlying log-logistic or Gompertz hazard is commonly found in the real world, we found that simulating under these hazards did not produce significantly worse

model misspecification (in terms of various tests' type I error and power) than for comparable Weibulls. This leads us to conclude that it is reasonable to proceed with a Weibull family assumption in the development and testing of new methods.

When there is no toxicity, or when parametric assumptions are met, the poly-k tests present an improvement over other available tests. However, we have seen that the tests' type I error and power are affected by toxicity when modeling assumptions are violated even within what is considered a reasonable range of Weibull settings. Due to strong parametric assumptions, none of the CA-based tests will perform well under all scenarios. These results motivate the development of a data-driven test that estimates the k parameter to avoid model misspecification in its application; Moon et al. have produced initial efforts attempting just that [26]. In Chapter 3 we describe an approach that appears to represent an improvement over this.

Chapter 3

AN MLE-BASED ESTIMATION OF K FOR THE POLY-K TREND TEST

3.1 Introduction

The poly-k test has been shown to have inflated type I error when its Weibull(k) modeling assumption is violated in the presence of treatment-associated intercurrent mortality. One way to improve upon this is to estimate the k shape parameter from the experimental data.

Moon et al. developed an estimator for k based on estimated lifetime cumulative tumor incidence rates [26]. Parameter k is estimated by equating two expressions for lifetime cumulative tumor incidence rate, and using this estimate in the poly-k test yields a "generalized poly-k test" that avoids the poly-k test's problems with treatment toxicity, by avoiding misspecification of k. Moon et al. report that their estimation method performs well for experiments that include 3 interim sacrifices, and recommend alternative methods when serial sacrifice information is not available. We note below that this estimator does not make full use of available data, which suggests that this method can be improved upon, either by reducing bias or by reducing reliance on interim sacrifice data.

Here, we propose another data-driven approach, using a maximum likelihood method to estimate k for a poly-$\hat{k}$ test. Through simulations of experiments with and without interim sacrifices, we examine the test's performance against existing methods, including the original poly-k for various levels of k, and Moon et al.'s method using estimated k.

3.1.1 Notation

Throughout this chapter, we will be referring to data and hazard expressions from a typical 2-year carcinogenicity experiment. Here are some key terms:

- treatment group $i = 0, 1, ..., I$; usually $I = 3$ and treatment $z_i \in (0, 0.25, 0.5, 1)$
- subject within i^{th} group $j = 1, ..., n_i$, where total sample size $n = \sum_i n_i$
- time to tumor s_{ij}
  time of death t_{ij}
  sacrifice indicator sac_{ij}
  tumor status $y_{ij} \in \{0, 1\}$
- $h_{z_i}^{DOC}(t) = L_{tx}^{z_i} h_{z_0}^{DOC}$ hazard for death from other causes, with $h_{z_0}^{DOC}(t) = \lambda^{k_D} k_D t^{k_D - 1}$;
  L_{tx} = treatment toxicity (treatment lethality)
- $h_{z_i}^{DFT|T}(t, s) = (L_{\text{tum}} - 1) h_{z_i}^{DOC}(t - s)$ conditional hazard for death from tumor, given that tumor occurred at time s; L_{tum} = tumor lethality parameter
- $h_{z_i}^{T}(t) = (1 + \eta z_i) h_{z_0}^{T}(t)$ tumor onset hazard in dose group z_i;
  $h_{z_0}^{T}(t) = \lambda^{k_T} k_T t^{k_T - 1}$ under the Weibull assumption;
  we use k and k_T interchangeably to refer to the tumor hazard shape parameter
- For each of the above we define the usual cumulative hazard and survival functions:
  $H_z(t) = \int_0^t h_z(u)\, du = \lambda^k t^k$
  $S_z(t) = \text{e}^{-H_z(t)} = 1 - F_z(t)$, where F is a CDF
- Dose effect on tumorigenicity $\eta = \frac{log(1-p_1)}{log(1-p_0)} - 1$, where $p_0 = F_{z_0}^T(t_{max})$ and $p_1 = F_{z_I}^T(t_{max})$ are the lifetime cumulative tumor incidence for the lowest and highest dose groups (assuming no intercurrent mortality)

3.1.2 Moon et al. k-estimation method

Acknowledging that the choice of k parameter for the poly-k test is somewhat arbitrary without additional guidelines, Moon et al. (2003) [26] proposed a method for estimating the value of k, to plug in to a "generalized poly-k test". To estimate k, they obtain two expressions for the lifetime cumulative tumor incidence rate (LCTIR). Assuming a Weibull(λ, k) tumor incidence rate function yields the first expression:

$$L = \prod_{\substack{i=0,\ldots,I \\ j=1,\ldots,n_i}} F^T(t_{max})^{y_{ij}} S^T(t_{ij})^{1-y_{ij}} = \prod_{i,j} \left[1 - \mathrm{e}^{-\lambda^k t_{max}^k}\right]^{y_{ij}} \left[\mathrm{e}^{-\lambda^k t_{ij}^k}\right]^{1-y_{ij}} \tag{3.1}$$

Equation 3.1 considers the cumulative probability of tumor at the terminal sacrifice time (t_{max}) for n animals, with tumor indicator y_{ij} and death time without tumor t_{ij}. Moon et al. treat (3.1) as though it were a likelihood. Differentiating with respect to λ, one can then solve in the usual way for the MLE $\hat{\lambda}(k)$. Plugging this back in to equation (3.1) yields, by the invariance property, the MLE $\widehat{\mathrm{LCTIR}}_1$, as shown in (3.2) below.

$$\widehat{\mathrm{LCTIR}}_1(k) = \prod_{i,j} \left[\frac{\sum_{i=1}^n y_{ij}}{\sum_{i=1}^n y_{ij} + \sum_{i=1}^n \alpha_{ij}(1-y_{ij})}\right]^{y_{ij}} \left[\frac{\sum_{i=1}^n \alpha_{ij}(1-y_{ij})}{\sum_{i=1}^n y_{ij} + \sum_{i=1}^n \alpha_{ij}(1-y_{ij})}\right]^{\alpha_{ij}(1-y_{ij})} \tag{3.2}$$

Note that (3.2) is a function of k through the time at risk adjustment α_{ij} described by Bailer & Portier for the poly-k test [3]; recall from section 1.2.2 that $\alpha_{ij} = (t_{ij}/t_{max})^k$ for non-tumor-bearing animals. Expression (3.2) can be evaluated as an empirical quantity, in terms of data from the experiment.

Separately, Moon et al. note that one may obtain a non-parametric discrete time model estimate of lifetime cumulative tumor incidence that is not a function of k. The estimate $\widehat{\mathrm{LCTIR}}_2$ (eqn (3.3)) is a conditional probability obtained using constrained numerical methods described by Kodell & Ahn [1, 19, 20]. Each MLE $\hat{\lambda}_m^T$ is an estimate of tumor incidence rate for the m^{th} sacrifice interval, i.e. $\lambda_m^T = P(T_T = m | T_D \geq m, T_T \geq m)$. Ahn & Kodell (1995) [1] estimated the quantities in $\hat{\lambda}_m^T$ by treatment group, whereas Moon et al. estimate

tumor incidence rate by pooling all animals in an interval.

$$\begin{aligned} \widehat{\mathrm{LCTIR}}_2 &= 1 - \prod_{m=1}^{M}(1-\hat{\lambda}_m^T) \\ \hat{\lambda}_m^T &= 1 - \left[(1-\hat{p}_m^A)(1-\hat{\lambda}_m^D) + (1-\hat{p}_m^D)\hat{\lambda}_m^D\right]/(1-\hat{p}_{m-1}^A) \end{aligned} \tag{3.3}$$

Adapting Ahn & Kodell's notation [1], let N_m^1 and N_m^2 be the number of natural deaths in interval m with and without tumors, and N_m^3 and N_m^4 be the number of sacrifices in interval m with and without tumors. The number of animals alive at the start of interval m is A_m and the total number of natural deaths is $N_m = N_m^1 + N_m^2$. Finally, the hazard function for time to sacrifice is denoted λ_m^S and the total number of sacrifices $S_m = N_m^3 + N_m^4$. Maximizing the likelihood in (3.4) yields estimates of the tumor prevalence for live animals $p_m^A = P(T_T \leq t_m | T_D > t_m)$, tumor prevalence for dead animals $p_m^D = P(T_T \leq t_m | T_D = t_m)$, and the discrete hazard rate for time to death $\lambda_m^D = P(T_D = t_m | T_D \geq t_m)$. These estimates are then substituted into (3.3) to obtain $\widehat{\mathrm{LCTIR}}_2$.

$$\begin{aligned} L \propto & \prod_{m=1}^{M}(p_m^D)^{N_m^1}(1-p_m^D)^{N_m^2}(p_m^A)^{N_m^3}(1-p_m^A)^{N_m^4}(\lambda_m^D)^{N_m}(1-\lambda_m^D)^{A_m-N_m} \\ & \cdot \prod_{m=2}^{M}(\lambda_{m-1}^S)^{S_{m-1}}(1-\lambda_{m-1}^S)^{A_m} \end{aligned} \tag{3.4}$$

$$\hat{p}_m^A = N_m^3/S_m$$

$$\hat{p}_m^D = N_m^1/N_m$$

$$\hat{\lambda}_m^D = N_m/A_m$$

Ahn & Kodell note that using these quantities can lead to negative values of $\hat{\lambda}_m^T$. To avoid algebraic intractability of maximizing under the constraint that $\hat{\lambda}_m^T \geq 0$, they propose an iterative algorithm for adjusting the estimates when the unconstrained optimization results in $\hat{\lambda}_m^T < 0$. Their method is described in detail in their original work. [1]

Setting $\widehat{\mathrm{LCTIR}}_1(k) = \widehat{\mathrm{LCTIR}}_2$, one can solve to obtain $\hat{k}$ directly (Moon et al. use a bisection method to obtain roots) and then use this estimate as the parameter in the usual

poly-k test. We will refer to Moon et al.'s generalized poly-k test as poly-$\hat{k}_{Moon}$.

In simulations with 3 interim sacrifices, with 50 animals per treatment group and 6 animals per group sacrificed at each interim timepoint, Moon et al. found that the poly-$\hat{k}_{Moon}$ test was robust to a variety of true underlying tumor hazards and had a type I error rate close to the poly-k test with correctly-specified k [26]. Notably, this method requires interim sacrifices, and the estimate of k is less biased in simulations with increasing numbers of sacrificed animals. Moon et al. examined how biased $\hat{k}$ was under varying combinations of sample size and number of animals sacrificed per interim sacrifice time. In all scenarios, the experiment had 3 interim sacrifice times. Under the (primarily rare tumor) conditions they simulated, when treatment group sample size was 50, increasing the number of animals sacrificed from 6 to 10 per sacrifice time reduced the bias of $\hat{k}$ for true $k \in (1.5, 3, 6)$; e.g. for true $k = 3$, mean $\hat{k}$ over 1000 data sets was 3.29 with 6 sacrificed animals, and 3.16 with 10 sacrificed animals. However, increasing the number sacrificed to 15 led to increased bias (e.g. under true $k = 3$, mean $\hat{k}$ was 3.32); this is likely due to insufficient effective sample size, as this represents a total of 45/50 animals interim-sacrificed over the whole experiment. For a per group sample size of 300, increasing the number of sacrifices (up to 90 animals per time point) always lowered bias. The authors also examined sample sizes of 500 and 1000 per treatment group and found that bias improved in particular for true $k \in (1.5, 6)$, and improved more with higher numbers of sacrificed animals (up to at least 90% of the sample); for true $k = 3$, these experimental designs were less optimal than a smaller sample size: for instance, with $N = 300$ and 90 sacrificed animals per interim time, mean $\hat{k}$ was 2.90, but with $N = 500$ and 150 sacrificed animals mean $\hat{k}$ was 2.73. Full results may be viewed in Moon et al. Table III [26]. We note that sample sizes greater than 100 per treatment group are extremely uncommon in these types of studies. For instance, the NTP 2-year study protocol has 50 animals per treatment group, sex, and species, with each sex-species combination typically analyzed separately [28].

3.1.3 Maximum likelihood estimation of k

The likelihood in expression (3.5), below, may be evaluated in terms of any parametric underlying hazard assumptions [23]. To apply this to estimating the k parameter for the Weibull-based poly-k test, we use Weibull hazards, with the elements that were defined more explicitly in Section 3.1.1. This results in a likelihood maximization problem in 7 parameters (k_T, λ_T, k_D, λ_D, L_{tum}, L_{tx}, η). Here, we set the treatment effect parameter $\eta = 0$ under H_0.

$$
\begin{aligned}
L \propto \prod_{i=1}^{I} \Bigg\{ \prod_{j=1}^{n_i} \Bigg[& \int_0^{t_{ij}} h_{z_i}^{T}(s) \left(h_{z_i}^{DOC}(t_{ij}) + h_{z_i}^{DFT|T}(t_{ij}, s) \right) \\
& \cdot \left(\mathrm{e}^{-\int_0^{tij} h_{z_i}^{DOC}(u)\,du - \int_0^{s} h_{z_i}^{T}(u)\,du - \int_s^{tij} h_{z_i}^{DFT|T}(u,s)\,du} \right) ds \Bigg]^{y_{ij}(1-sac_{ij})} \\
& \cdot \left[h_{z_i}^{DOC}(t_{ij}) \left(\mathrm{e}^{-\int_0^{t_{ij}} \left(h_{z_i}^{DOC}(u) + h_{z_i}^{T}(u) \right) du} \right) \right]^{(1-y_{ij})(1-sac_{ij})} \\
& \cdot \left[\int_0^{t_{ij}} h_{z_i}^{T}(s) \left(\mathrm{e}^{-\int_0^{tij} h_{z_i}^{DOC}(u)\,du - \int_0^{s} h_{z_i}^{T}(u)\,du - \int_s^{tij} h_{z_i}^{DFT|T}(u,s)\,du} \right) ds \right]^{y_{ij} sac_{ij}} \\
& \cdot \left[\mathrm{e}^{-\int_0^{t_{ij}} \left(h_{z_i}^{DOC}(u) + h_{z_i}^{T}(u) \right) du} \right]^{(1-y_{ij}) sac_{ij}} \Bigg\}
\end{aligned}
\tag{3.5}
$$

To bound the parameter estimates, we reparametrize (3.5), with $\alpha_T = log(k_T)$, $\alpha_D = log(k_D)$, $\beta_T = log(k_T \lambda_T^{k_T})$, $\beta_D = log(k_D \lambda_D^{k_D})$, $\phi = log(L_{tum} - 1)$, and $\tau = log(L_{tx})$.

To maximize the log likelihood, we use the Broyden, Fletcher, Goldfarb, Shanno (BFGS) algorithm [7, 8, 15, 16, 39], a quasi-Newton optimization method which we found to perform well in terms of convergence and to have faster speed of computation than other methods. In the rare event that BFGS is unable to estimate the gradient, the optimizer is able to use the slower Nelder-Mead simplex algorithm [27] to obtain an estimate; this has occurred in our simulations due to sparse data, in the case of rare tumors. Once MLE's are obtained, the estimate $\hat{k_T} = \mathrm{e}^{\hat{\alpha_T}}$ is readily calculated and used to conduct the poly-$\hat{k}_{MLE}$ test.

3.2 Simulation Study

We simulate 2-year (104 week) murine terminal sacrifice experiments with 60 animals per treatment arm and 4 dose groups (doses: 0 (control), 0.25, 0.5, and 1). Each scenario is run with 5,000 (under H_a) or 10,000 (under H_0) replicated data sets.

Weibull hazards ($h(t) = \lambda^k k t^{k-1}$) are widely used to model time-to-failure data, and this model is flexible enough to accommodate a wide variety of scenarios (see Figure 3.1 (left panel) and Chapter 2). For the primary set of simulations, times to tumor were modeled using Weibull hazards with $k_T \in (1, 1.5, 3, 6)$; the range of simulated k_T includes the values examined by Moon et al. [26]. Parameter λ_T is determined by our choice of study-end cumulative tumor incidence in the control group, absent intercurrent mortality: $p_0 = F^T(s = 104|k_T, \lambda_T) = 0.15$. A subset of simulations was also conducted using $p_0 = 0.05$, to assess how tests' performance might differ when tumors are rarer.

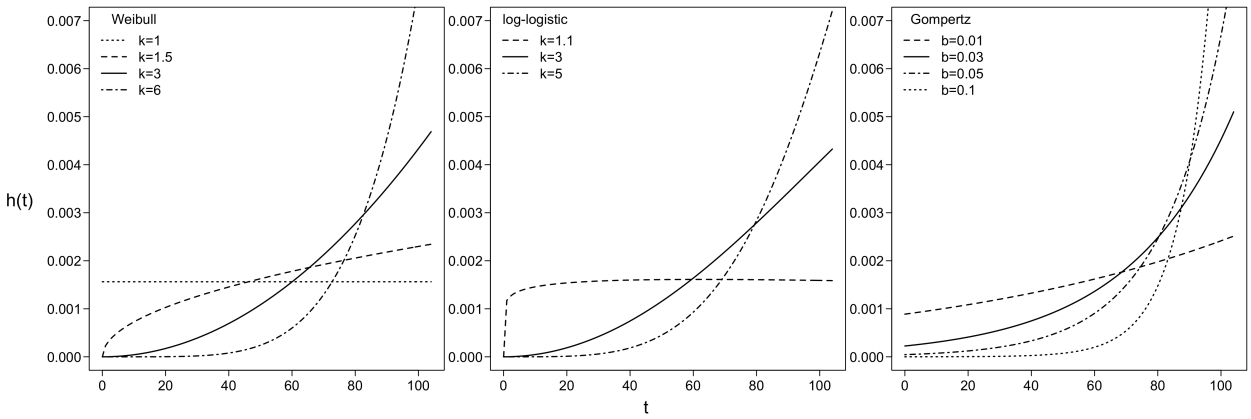


Figure 3.1: Tumor hazards used in simulation studies

Since we implement the log-likelihood assuming that the underlying tumor hazards are parametrized as a Weibull, we wish to assess whether good performance of the MLE approach is dependent on the true tumor hazard being a Weibull. A separate set of simulations was therefore also conducted using Gompertz ($b \in (0.01, 0.03, 0.05, 0.1)$) and log-logistic ($k \in (1.1, 3, 5)$) tumor hazards. (See Figure 3.1 to compare the simulated tumor hazard

shapes.) These settings include some examples with very different shapes from the usual range of Weibulls: Gompertz($b = 0.1$) puts almost all tumors at the end of the lifetime experiment and with log-logistic ($k = 1.1$) tumor hazard increases and then decreases over the course of 104 weeks. As with the Weibull simulations, the second parameter for log-logistic and Gompertz was determined by fixing $p_0 = 0.15$.

We vary tumor lethality - the extra risk of death due to presence of tumor - as being either $L_{tum} = 1$ (no lethality) or $L_{tum} = 1.5$. Treatment toxicity - a dose-associated mortality modifier - is simulated as either $L_{tx} = 1$ (no toxicity) or $L_{tx} = 1.8$. Times to death are modeled using Weibull hazards: the competing risk death hazard h^{DOC} is Weibull($k = 2, \lambda = 0.007$), with a control group study-end competing risks survival probability $S^{DOC}(t = 104, z = 0) = 0.6$. (We note that the simulated $L_{tx} = 1.8$ corresponds to a by-group competing risks survival probability of 0.6, 0.55, 0.5, and 0.4, respectively for the lowest through highest dose groups.) Under H_a, the cumulative tumor probability for the highest dose group is twice that of the control group: $p_1 = 2p_0$. For comparison with rarer cancers, a subset of these simulations (using $L_{tum} = 1$ and $L_{tx} = 1.8$) have also been conducted under $p_0 = 0.05$.

Methods used for generating times to tumor and times to death (from tumor or from other causes), and for applying lethality and toxicity effects, were described earlier in section 2.3.

Experiments with and without interim sacrifices are considered, with the number of interim sacrifices ranging from 0 to 4. See Table 3.1 for details on the experimental settings that were included across these simulations. At each interim time point, n animals in each treatment group are randomly selected for sacrifice. For each serial sacrifice design, n is chosen so that the total number of animals sacrificed at interim times over the experiment is comparable to the number under other sacrifice designs. For example, with 1 sacrifice, 10 animals were selected per treatment group, for a total of 10 sacrificed; whereas with 2

interim sacrifices, 5 were selected at each time point to maintain a total of 10 sacrificed.

For an illustration of how the shape of the data is impacted by the simulation design, in particular by the true value of k_T, the application of interim sacrifices, and then number of animals sacrificed at each interim time, please refer to the plots in Appendix C.1.

Table 3.1: Simulation Experimental Design Settings: Serial Sacrifices

# interim sacrifices	Sacrifice times	N (per tx group)	n per interim sacrifice	Total sacrificed (interim)
0	104	50 †	—	—
0	104	60	—	—
1	52, 104	60	10	10
2*	52, 78, 104	60	5	10
3	52, 78, 92, 104 ‡	60	4	12
4*	52, 65, 78, 91, 104	60	3	12

† With 1 interim sacrifice, NTP uses N=60 per group. Varying sample size here allows us to separate the effect of additional sample size from effect of having interim sacrifices.
‡ These sacrifice intervals are the NTP standard, according to Dinse [11].
* Equally spaced sacrifices. (3 interim sacrifices uses "more near the end" strategy.)

We examine the average $\hat{k}$ estimates from both the Moon et al. procedure and our maximum likelihood based method. We also compare type I error and power (calibrated to 0.05 size) for the 1-sided versions of the following tests:

- poly-$\hat{k}_{MLE}$, our newly proposed test
- poly-$\hat{k}_{Moon}$, the estimated-k test proposed by Moon et al.
- poly-k test with Bieler-Williams variance [3, 5] and pre-specified $k \in (1.5, 3, 6)$
- Cochran Armitage test [2]
- log rank test for trend [40]
- Hoel-Walburg (Mantel-Haenszel) test with intervals based on sacrifice times (minimum of 4 intervals), treating sacrificed animals separately from natural deaths
- logistic score test (adjusting for continuous time with linear and quadratic terms) [12]

3.3 Results

3.3.1 MLE method convergence

Overall, the MLE k-estimation method using BFGS optimization proved to be a stable algorithm, although there were cases for which BFGS was not able to produce an estimate and the algorithm had to utilize Nelder-Mead optimization. In Weibull simulations with moderate cancer rate data ($p_0 = 0.15$), the BFGS optimization never failed. For low cancer rate data ($p_0 = 0.05$), there were some missing MLE values where BFGS could not run due to sparse data. This occurred in 0.11% of $800,000$ H_0 simulations and only 24 times out of $400,000$ H_a simulations, where more tumors could be observed due to the simulated treatment effect. Algorithm failure was more common in the presence of tumor lethality and treatment toxicity, as well as for higher true k_T. That higher k_T will be more likely to result in insufficient numbers of tumors in the data is readily seen by inspecting the plots provided in Appendix C.1 (Figures C.1.2, C.1.3, C.1.4, C.1.5), and noticing that early accrual of animals observed with tumor is much lower as k_T increases.

When data were simulated using log-logistic tumor onset hazards (with moderate cancer rate), the BFGS optimization never failed. With the exception of $k = 1.1$, the shapes of the log-logistic hazard curves were relatively close to the usual range of Weibull hazards assumed for these sort of data, so this result is unsurprising (see Figure 3.1). Also, with log-logistic $k = 1.1$, the hazard of tumor is not especially low for most of the experiment, so it does not produce the sort of sparse tumor data that would result in problems evaluating and optimizing the likelihood.

When data were simulated using Gompertz tumor hazards, there were also some missing MLE values due to sparse data. This occurred about 0.01% of the time under H_0 (57/480000 simulated data sets) and less frequently under H_a (7/240000 data sets). This occurred most frequently for shape parameter $b = 0.1$, which makes sense since that was the most extreme

shape considered, with almost all tumors appearing only near or at the end of the study. Cases that did and did not fail might have similar numbers of tumor bearing animals by the end of the experiment, but those that failed tended to have no tumor bearing animals with death times (natural or sacrificed) prior to 103 weeks. This kind of data sparsity reduces the available information for estimating the gradient vector required by BFGS optimization.

When the BFGS optimization failed to run, the MLE algorithm switched to using the Nelder-Mead simplex method. A handful of these Nelder-Mead runs failed to converge due to a degenerate simplex: these are indicated in Tables C.5 and C.6 where prop_converged is < 1. In those cases, the most recent value for $\hat{k_T}$ was used. We also implemented a fall-back routine for non-convergence cases: the test could optionally be run instead as a standard poly-3 test or as a poly-$\hat{k}_{Moon}$. The difference these versions of the algorithm made to type I error and power were not discernable up to 4 decimal places, so to avoid redundancy we omit the fall-back results here.

3.3.2 Simulation Results: Bias

As noted earlier, the k estimator by Moon et al. requires one or more interim sacrifices; where there is only a terminal sacrifice, their algorithm outputs $\hat{k}$ values close to zero, effectively turning the $poly - \hat{k}_{Moon}$ test into the Cochran-Armitage test. Our MLE-based method is able to produce k estimates even using experiments without interim sacrifices, although for moderate cancer rate data ($p_0 = 0.15$), the MLE's of k are more biased than Moon's estimate when there are 3 interim sacrifices; this is true regardless of whether the treatment is toxic or tumors are lethal. (Figure 3.2.) That the MLE is a consistent estimator is suggested empirically by a reduction in bias under simulations which increase the sample size (some results shown in Chapter 4).

Among the different serial sacrifice designs, in our simulated experiments, the Moon estimator is typically least biased when there are 4 sacrifices (3 interim). With fewer sacrifices

it underestimates k, and with 5 sacrifices, it overestimates. It is worth noting that Moon et al.'s paper describing this method focused on the 4 sacrifice design in particular [26]. By contrast, the MLE-based estimator tends to overestimate k at lower numbers of sacrifices, becoming slightly less biased with more sacrifice times (Figure 3.2). One exception is that the MLE becomes more biased with 1 interim sacrificed (compared to none) when true $k = 6$, and then less biased for increasing interim sacrifices. With $k = 6$, more tumors are later onset, so a single sacrifice likely removes animals from the experiment too early.

Both the MLE and Moon estimator tend to produce lower k estimates under H_a than under H_0 (Figure 3.3). With higher numbers of sacrifices, the Moon et al. estimate is sometimes less biased, and is usually more variable than the correspoding MLE estimate. For example, in simulated data with treatment toxicity but no tumor lethality, when there were 4 total sacrifices, and assuming H_0 true and a true k_T of 3 (see Table C.2 in Appendix C.2), both the Moon estimate (mean $\hat{k} = 3.44$) and the MLE (mean $\hat{k} = 3.74$) overestimated the parameter on average, but Moon came closer. Here, Moon was more variable than the MLE (SD 1.95 versus 1.67). Under H_a with the same simulation settings, both estimates were less biased and less variable than under H_0; this tended to be true in general. In this case, under H_a, mean $\hat{k}$ using the Moon method was 2.97 (SD 1.52) whereas the mean MLE was 3.31 (SD 1.20).

For rarer cancers, both methods produce higher estimates than under the analagous settings with more common cancers (Figures 3.4 and 3.5). Both methods tended to be more variable for rarer cancers (compare, for example, SDs between Tables C.2 and C.5 in Appendix C.2), which makes sense since fewer tumor events represent a scarcity of information for producing estimates.

Another way to compare these two estimators is to examine their mean squared error (MSE) for each simulation setting (Figures 3.6, 3.7). The two tests' behavior was very similar regardless of toxicity or lethality setting. With a single, terminal, sacrifice time, the Moon

Figure 3.2: Mean $\hat{k}_T$ when H_0 true and $p_0 = 0.15$; varying k_T, lethality, toxicity (Weibull tumor hazard)

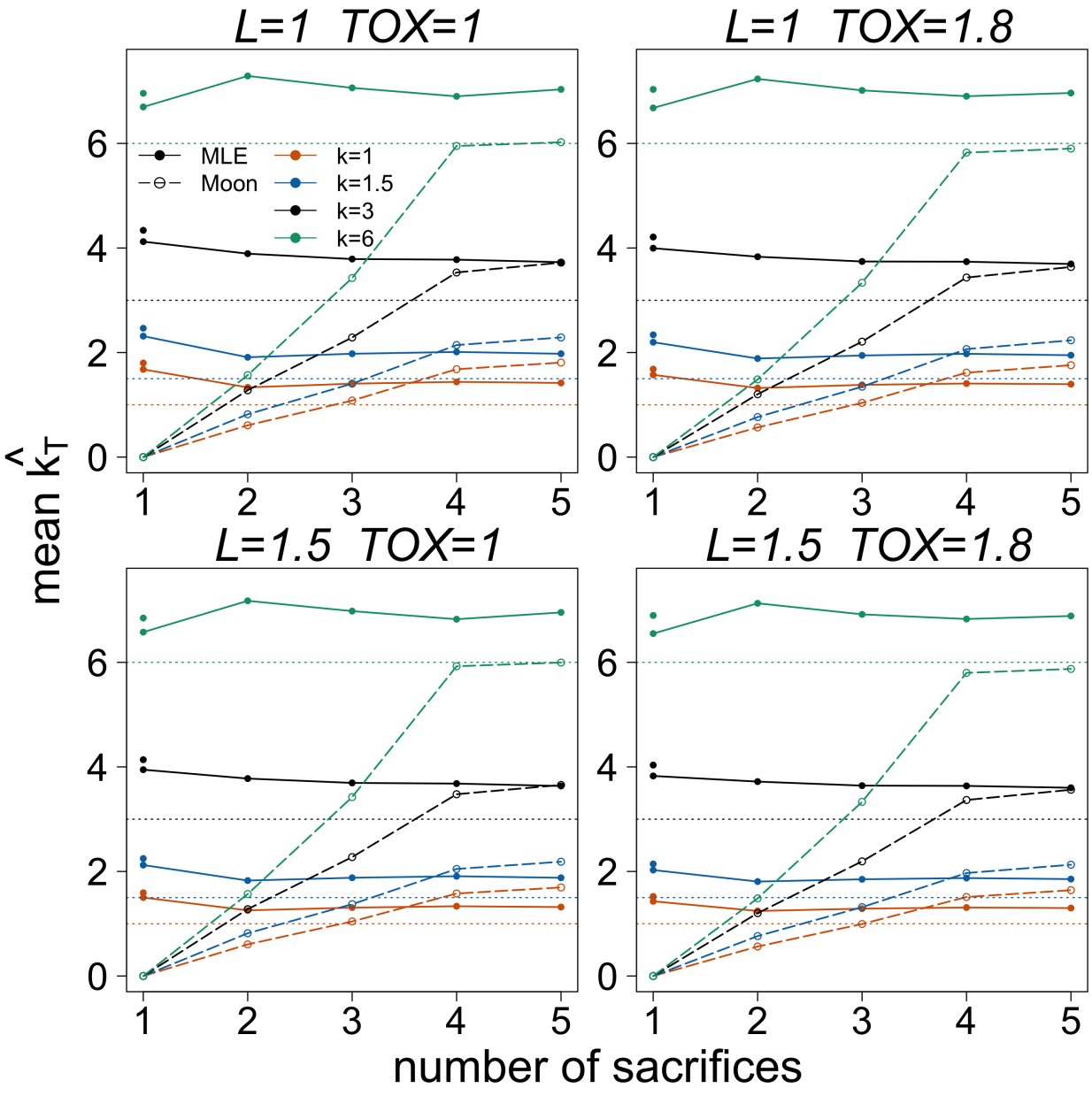


Figure 3.3: Mean $\hat{k}_T$ when H_a true and $p_0 = 0.15$; varying k_T, lethality, toxicity (Weibull tumor hazard)

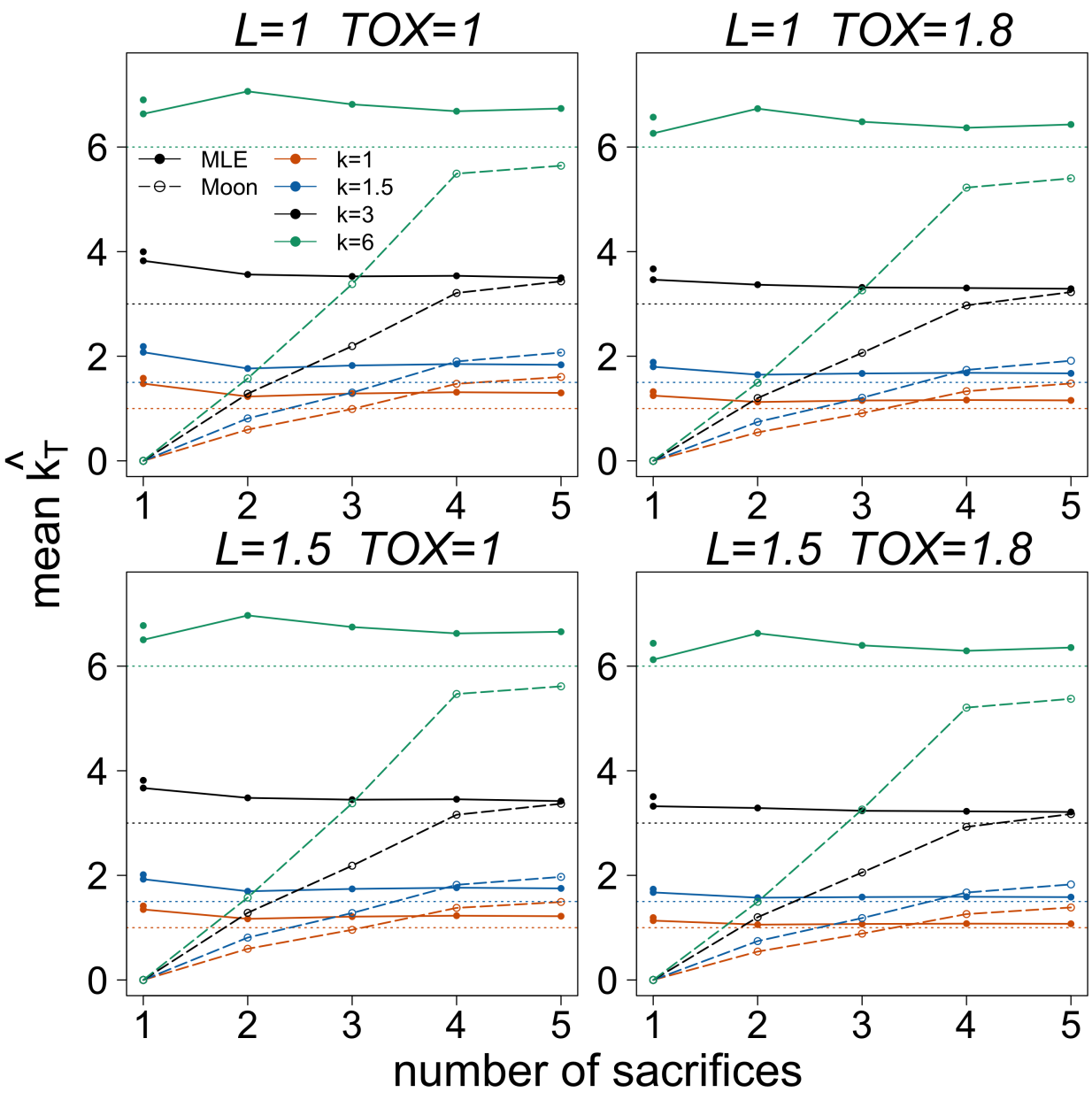


Figure 3.4: Mean $\hat{k}_T$ when H_0 true and $p_0 = 0.05$; varying k_T, lethality, toxicity (Weibull tumor hazard)

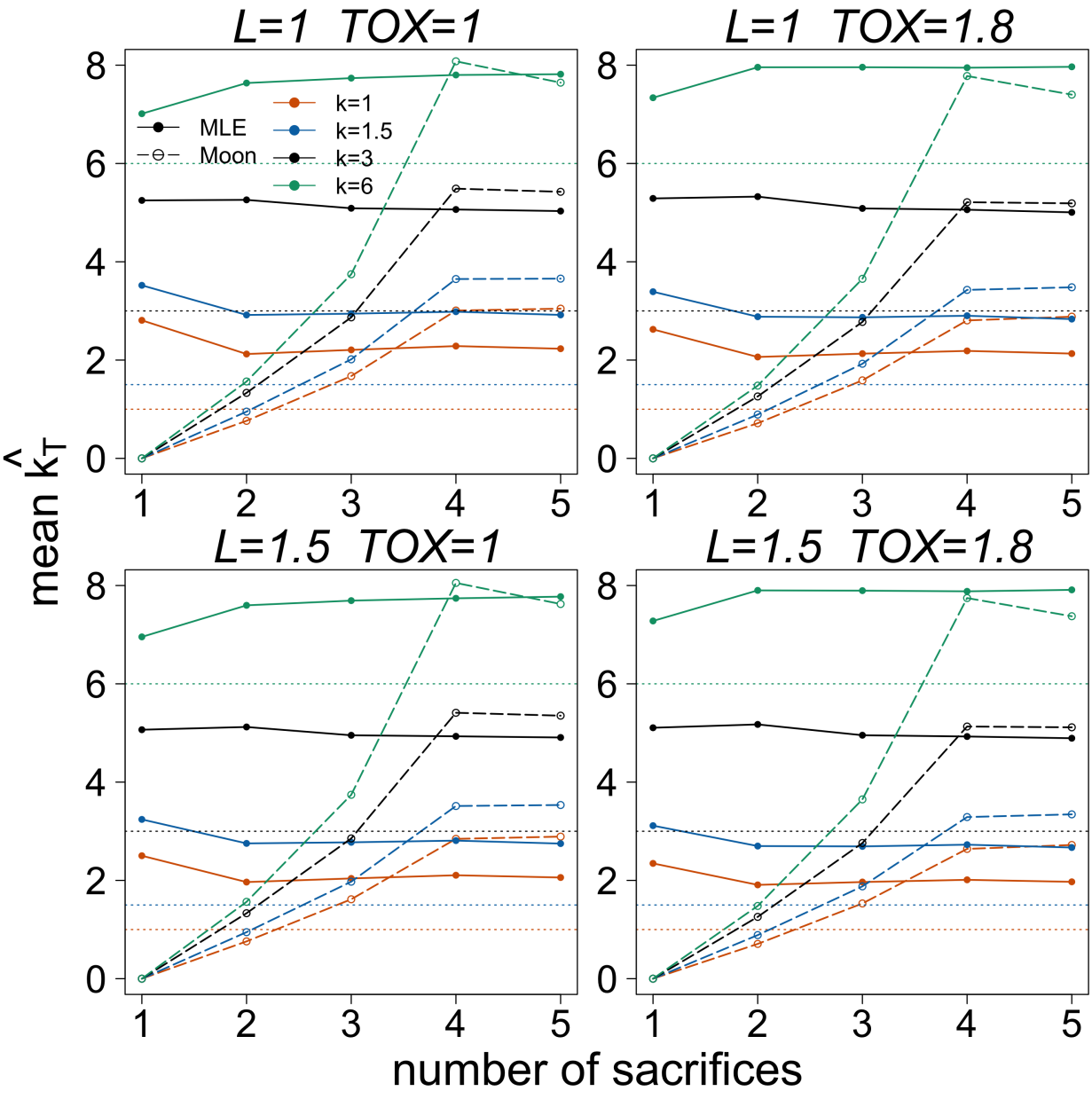


Figure 3.5: Mean $\hat{k}_T$ when H_a true and $p_0 = 0.05$; varying k_T, lethality, toxicity (Weibull tumor hazard)

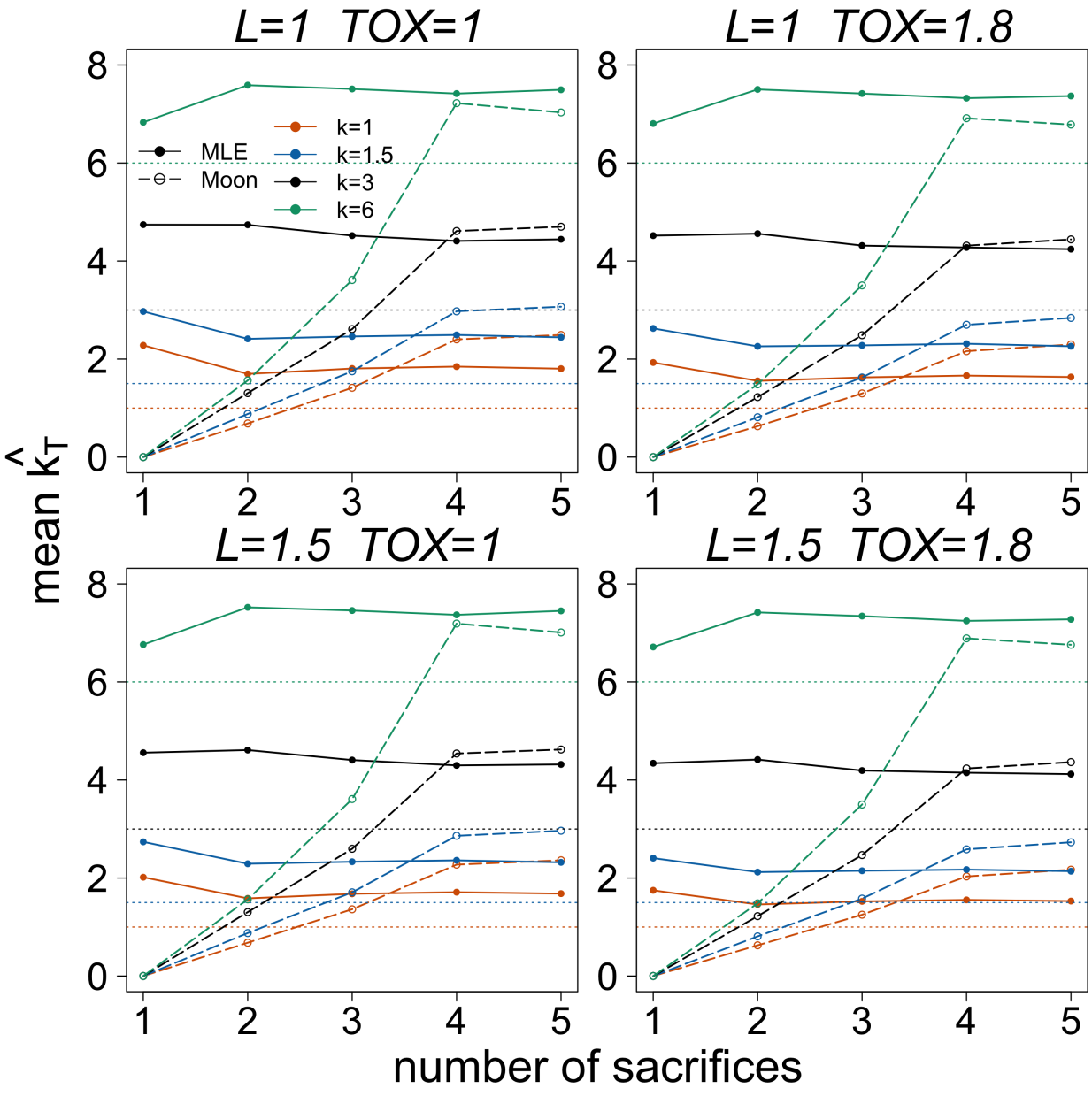


Figure 3.6: MSE of $\hat{k}_T$ when H_0 true and $p_0 = 0.15$; varying k_T, lethality, toxicity (Weibull tumor hazard)

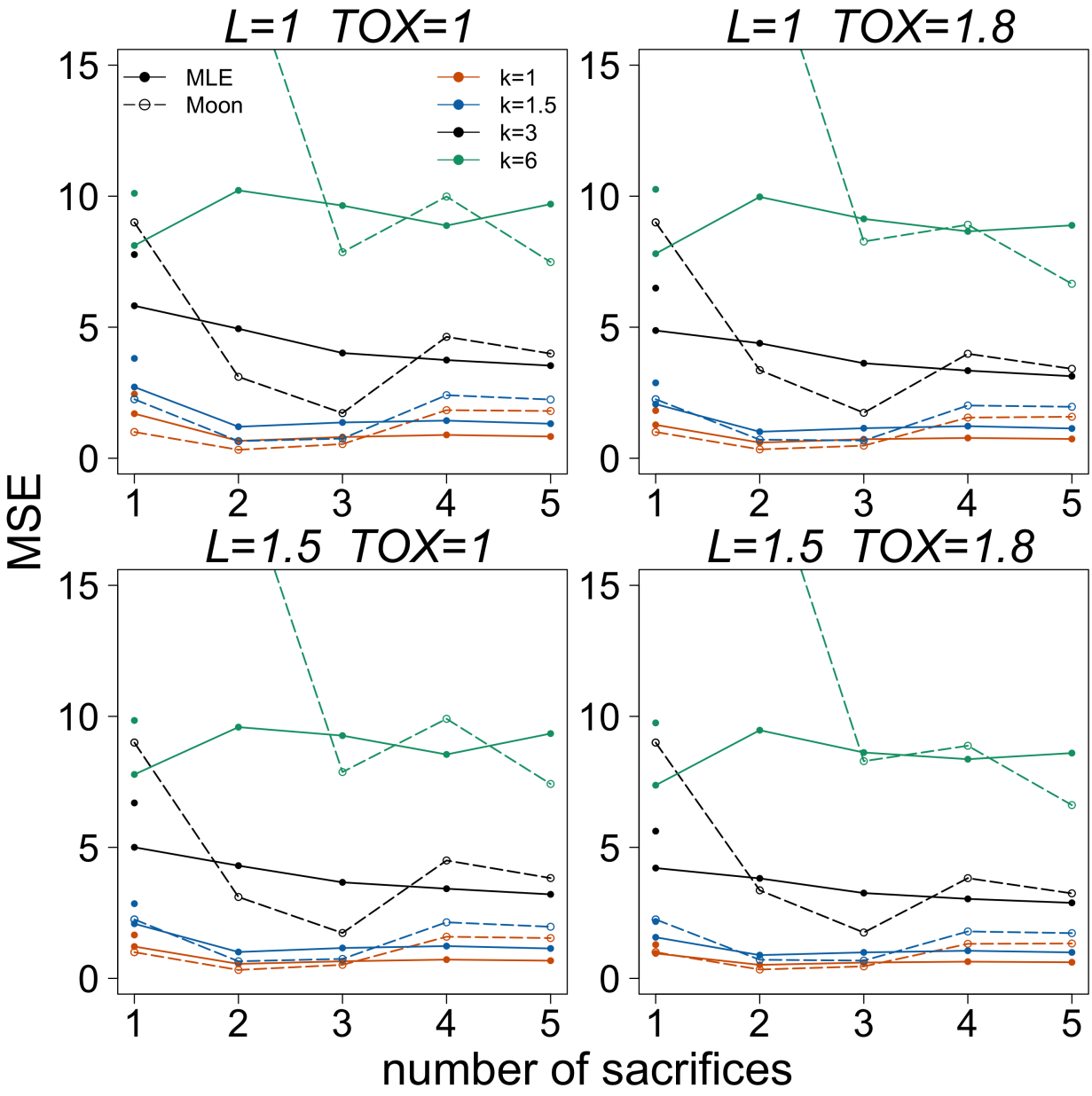


Figure 3.7: MSE of $\hat{k}_T$ when H_a true and $p_0 = 0.15$; varying k_T, lethality, toxicity (Weibull tumor hazard)

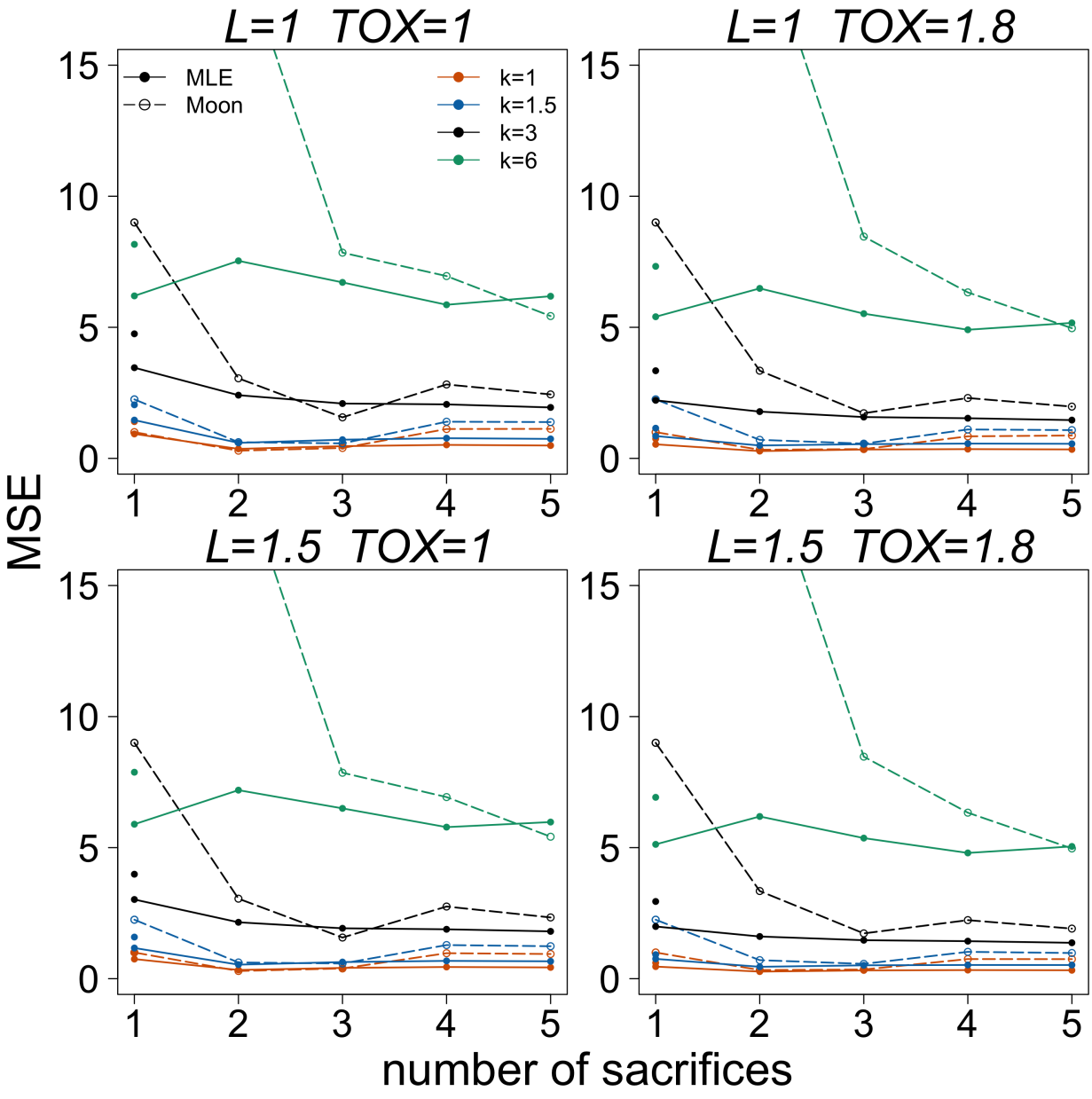


Figure 3.8: MSE of $\hat{k}_T$ when H_0 true and $p_0 = 0.05$; varying k_T, lethality, toxicity (Weibull tumor hazard)

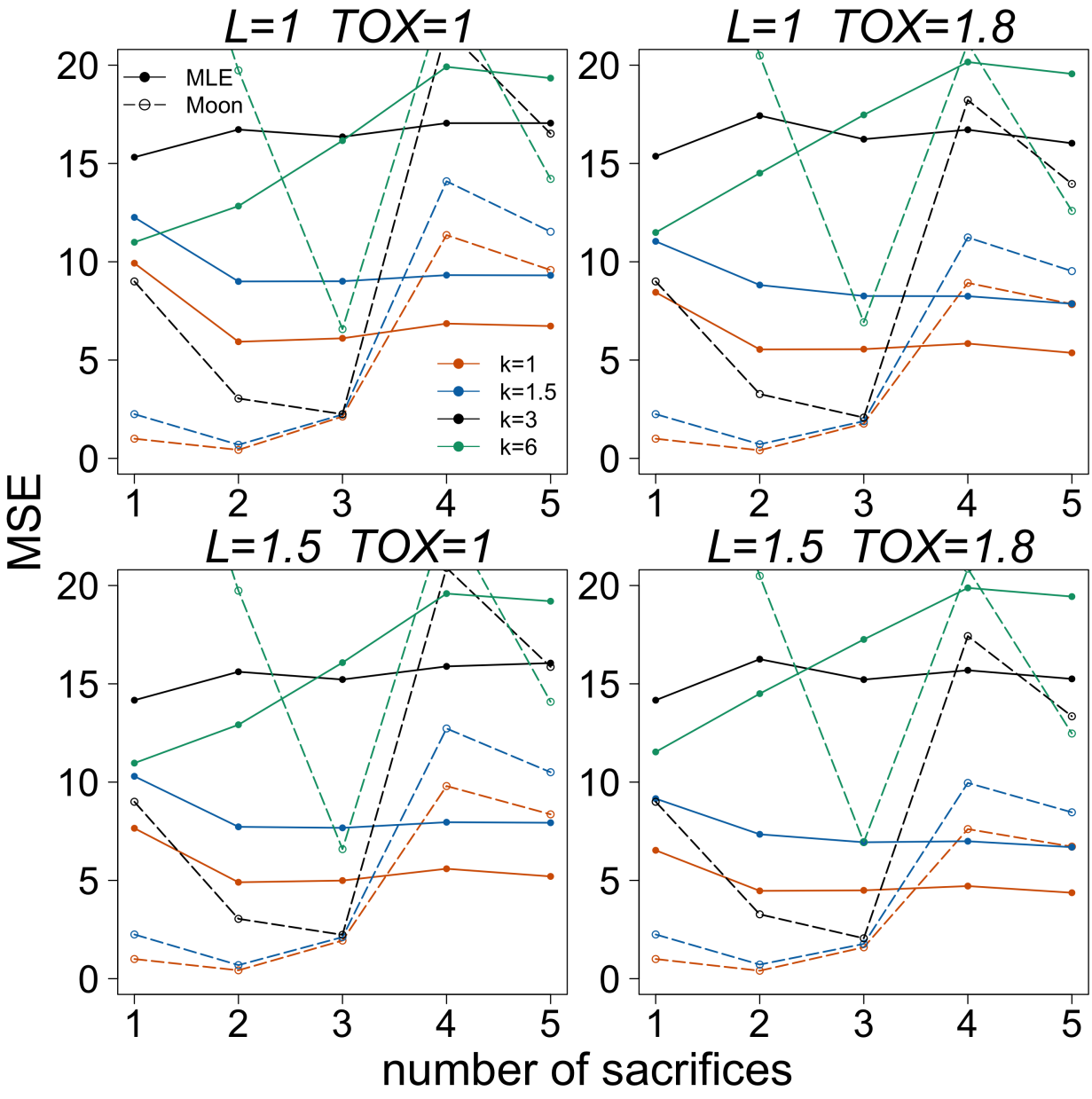
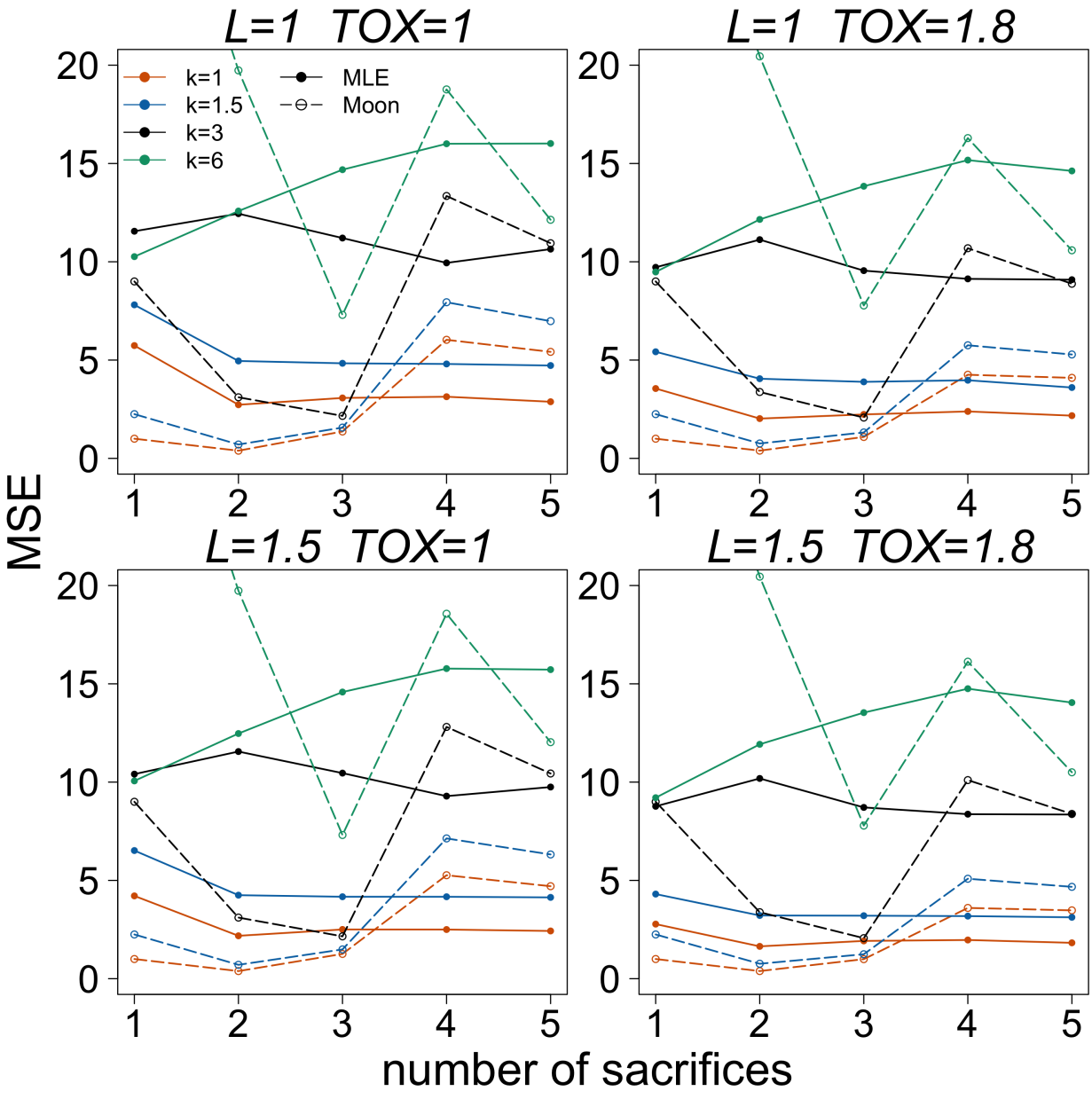


Figure 3.9: MSE of $\hat{k}_T$ when H_a true and $p_0 = 0.05$; varying k_T, lethality, toxicity (Weibull tumor hazard)



estimator had higher MSE than the MLE when true $k \in (3, 6)$, driven by its extreme bias in these cases. For 2 to 4 sacrifices, the MLE tended to have higher MSE despite tending to have a lower bias than $\hat{k}_{Moon}$; we noted earlier that the MLE was more variable in this mid-range for numbers of sacrifices. For 4 and 5 sacrifices, the Moon estimator's increasing variability results in its MSE falling above the MSE of $\hat{k}_{MLE}$, despite it being less biased at these settings. Overall, the MSE for the MLE decreases somewhat with increasing numbers of sacrifice times; the volatility in the comparison of the two estimators is driven by a somewhat U-shaped pattern to the Moon estimator's MSE over numbers of sacrifices. The relationship between the MSE's of $\hat{k}_{MLE}$ and $\hat{k}_{Moon}$ is similar when cancers are lower incidence (Figures 3.8, 3.9), although all MSE's are somewhat higher in that case. In the results presented below (section 3.3.3), it can be noted that the performance of the poly-$\hat{k}$ tests (in particular type I error) relates more closely with how far $\hat{k}$ is from the truth than it does to the MSE.

All numeric results for the figures discussed above are available in Appendix C.2, Tables C.1, C.2, C.3, C.4, and C.5.

For simulated data where tumor onset was simulated using non-Weibull hazards, it is not sensible to assess bias in estimation of the Weibull k_T parameter. However, in these cases we may consider how closely a Weibull hazard with the estimated $\hat{k}_{MLE}$ approximates the true underlying (log-logistic or Gompertz) hazard shape. In each panel of Figure 3.10, a mean Weibull k parameter was estimated from data sets generated using the indicated log-logistic or Gompertz hazard; these particular examples assumed treatment toxicity but no tumor lethality and used the $\hat{k}_{MLE}$'s obtained from experiments with 1 interim sacrifice. Our results show that the Weibull can often closely approximate these non-Weibull hazards, but with some notable exceptions.

The log-logistic with $k = 1.1$ is decidedly non-Weibull in shape, with a steep initial rise in $h(t)$ followed by a gradual rise and then decline over the course of the 104 week experiment. The approximated Weibull with $k = 1.37$ is also a poor fit for the concave-up shape of the

Gompertz with $b = 0.01$. We see in the next section that these departures for the most part do not impact test performance when the true underlying tumor hazards are these non-Weibulls.

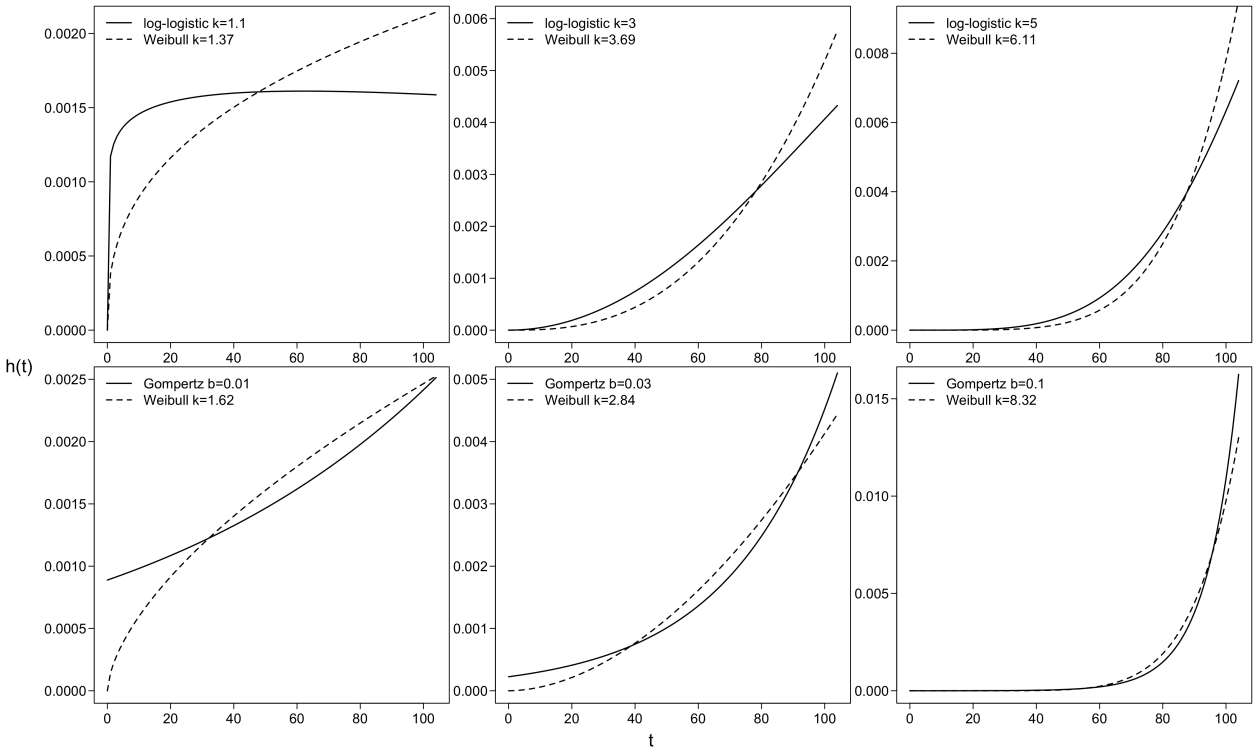


Figure 3.10: True tumor hazard (solid lines) for data generation versus Weibull with k parameter estimated as an average over many data sets

3.3.3 Simulation Results: Type I Error and Power

Type I error and power results reflect what was seen regarding estimation bias betwen these two methods, as well as what we have seen previously for the poly-k family of tests.

In the results for no toxicity in Figure 3.11 (left 2 columns), there is substantial overlap in type I error among the 3 choices of fixed poly-k test, along with the two poly-$\hat{k}$ methods, across all serial sacrifice designs considered. Since there is no treatment toxicity, the poly-k test performs consistently for any reasonable choice of parameter k. As we saw previously,

the poly-k family of tests also do not quite achieve nominal alpha level at this sample size. The comparison tests included in this plot (CA, logrank, logistic score, and MH) do reject at nominal alpha (0.05). Number of sacrifices does not affect type I error for any of the tests in the absence of toxicity, except for the case of no interim sacrifices. Comparing across results with $N = 60$ per group (all points joined by lines in the plots), the experiments with no sample taken for interim sacrifice have a sample size advantage, and tend to have slightly higher level than experiments with interim sacrifices. These plots also include the $N = 50$ single-sacrifice experiment results, as stand-alone dots, for comparison. These tend to have slightly lower level than the $N = 60$ experiments with 10 animals taken for interim sacrifice; this is unsurprising, since the sacrifices contribute information and do not simply reduce the sample size.

When toxicity is introduced (right two columns of Figure 3.11), specification of the k parameter in the poly-k tests becomes important to type I error; as discussed in Chapter 1 (section 1.3) and seen in simulations in Chapter 2, model misspecification can lead to conservative or anticonservative test behavior, depending on the direction of the misspecification. For each choice of true k, the poly-$\hat{k}_{MLE}$ type I error follows the poly-k line (with pre-specified k) whose k parameter happens to agree with the truth. For instance, when true $k = 3$, the poly-$\hat{k}_{MLE}$ and poly-3 type I errors are very close; when true $k = 6$, the MLE-based test type I error is very similar to that of the poly-6 test. By comparison, the poly-$\hat{k}_{Moon}$ test is very conservative for experiments with low numbers of sacrifices, and only matches the level of poly-$\hat{k}_{MLE}$ when the total number of sacrifices is greater than 3. This pattern persists for lower incidence cancers (Figure 3.13).

The classical tests we included for comparison were the logrank, MH, logistic score, and CA tests. Their behavior in these simulations is consistent with what has been seen previously. For instance, the logrank has inflated type I error in the presence of toxicity. This behavior is more extreme for lower values of true underlying shape parameter; under

Figure 3.11: Type I error: $p_0 = 0.15$, varying k_T, lethality, toxicity (Weibull tumor hazard)

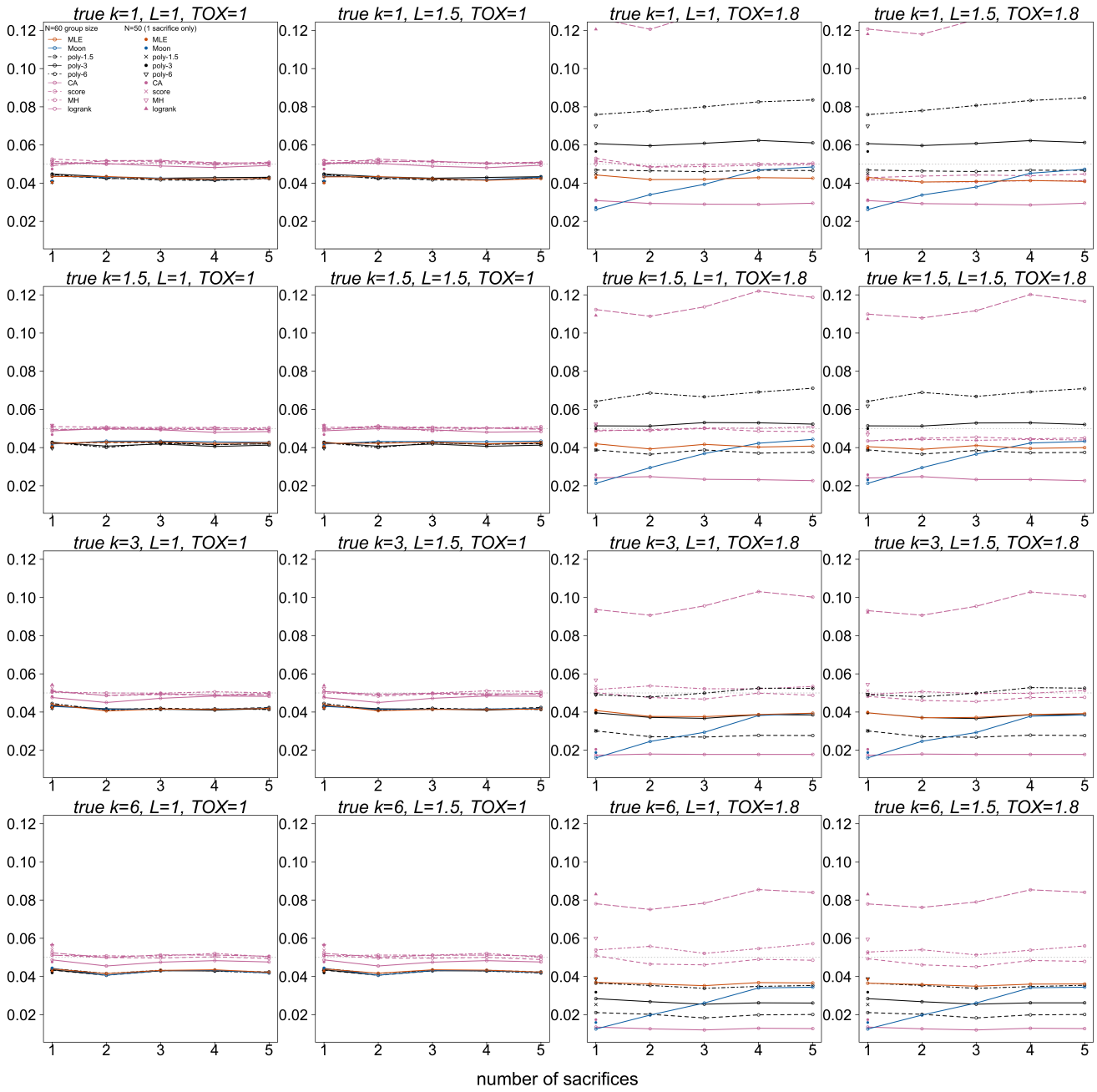


Figure 3.12: Power: $p_0 = 0.15$, varying k_T, lethality, toxicity (Weibull tumor hazard)

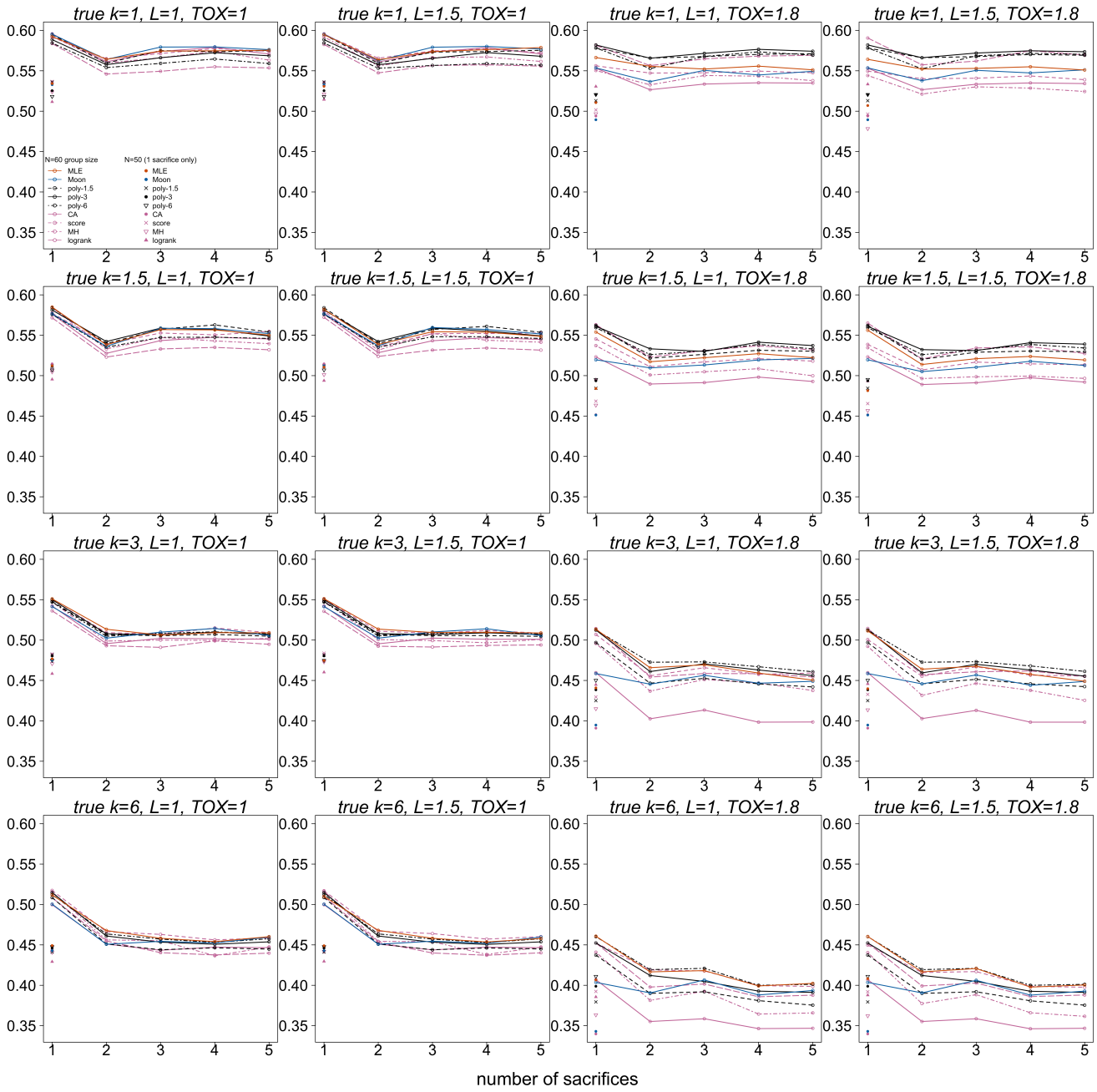


Figure 3.13: Type I error: $p_0 = 0.05$, varying k_T, lethality, toxicity (Weibull tumor hazard)

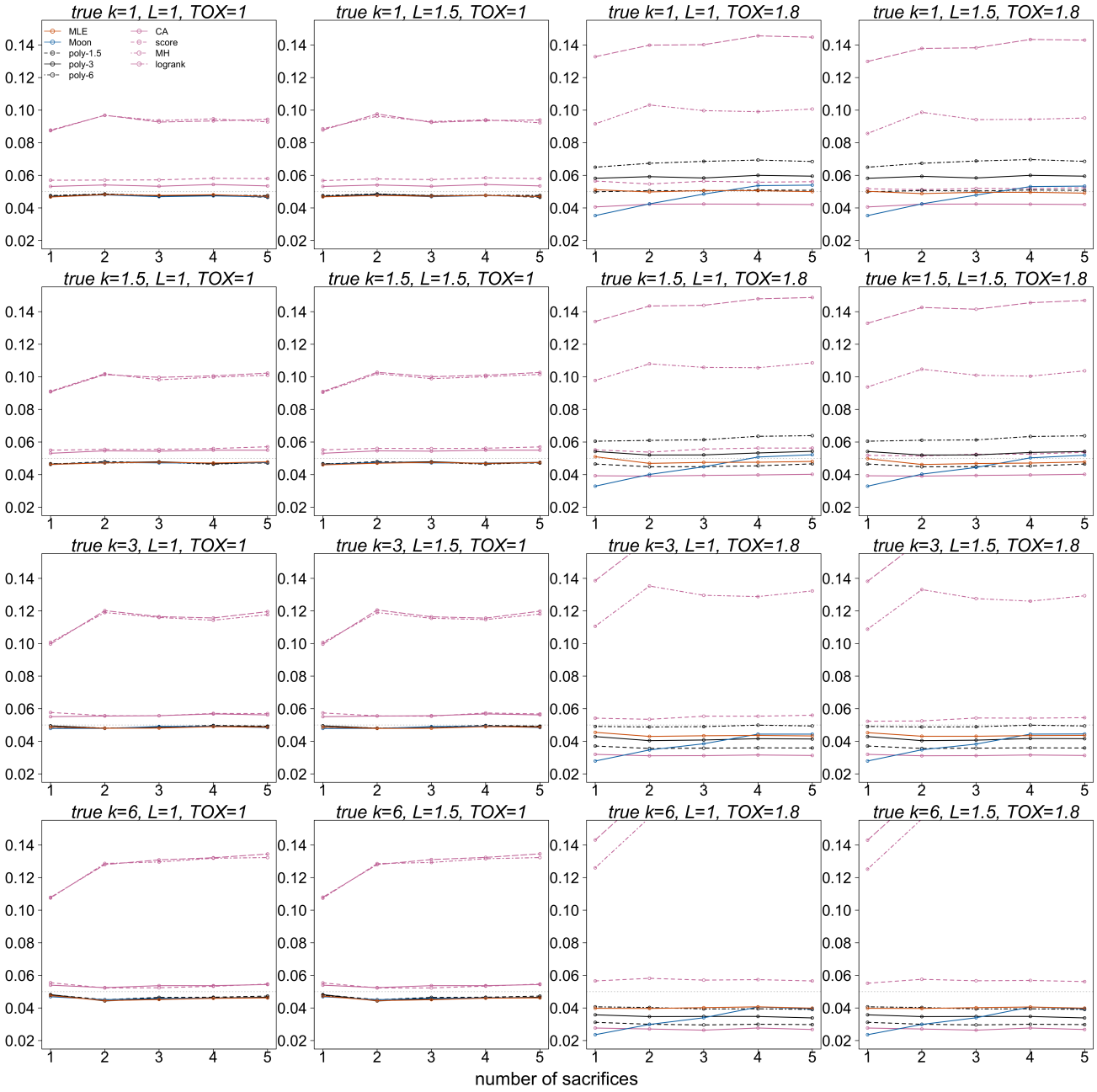


Figure 3.14: Power: $p_0 = 0.05$, varying k_T, lethality, toxicity (Weibull tumor hazard)

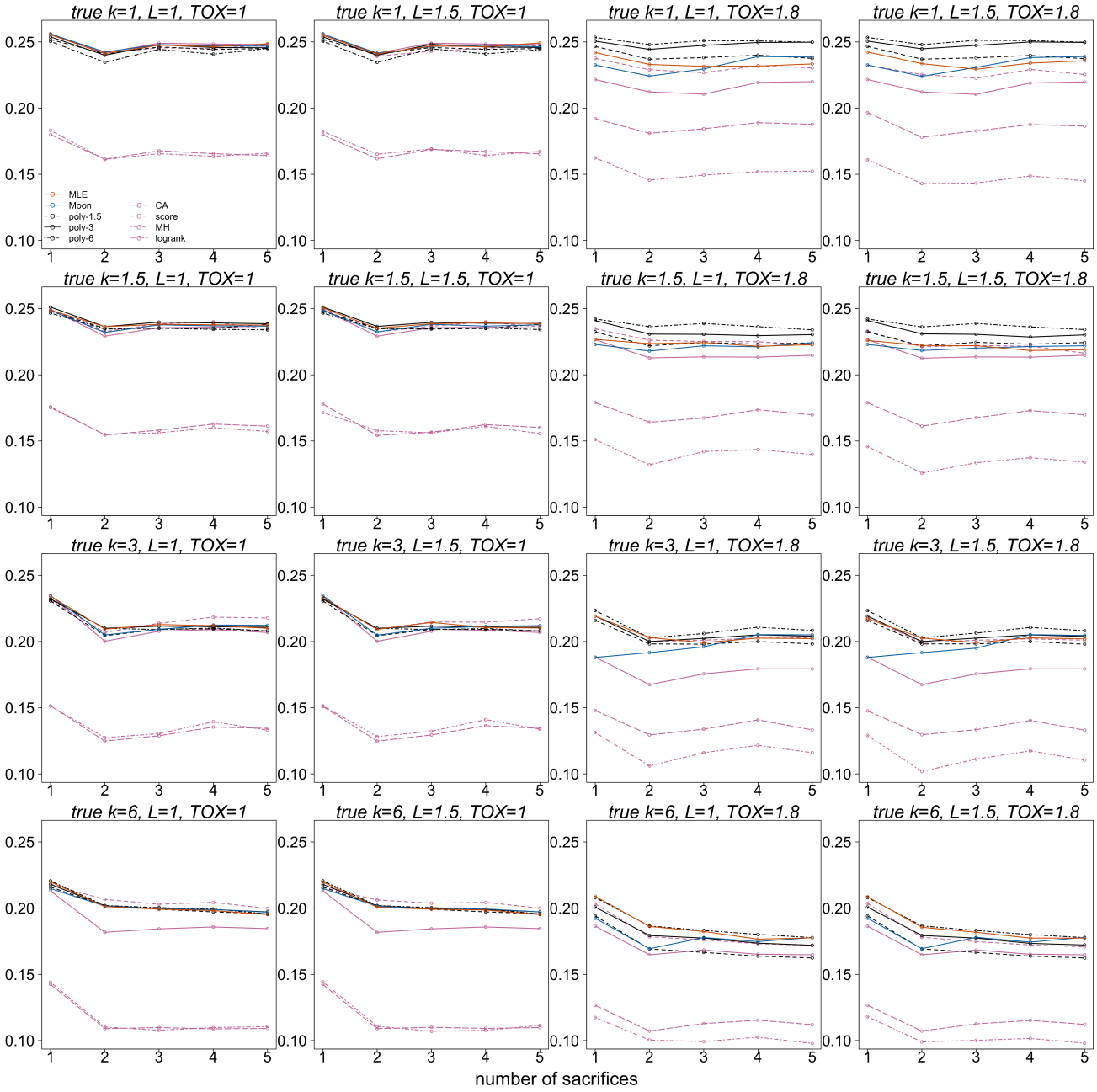


Figure 3.15: Type I error: $p_0 = 0.15$, varying k_T, lethality, toxicity (log-logistic tumor hazard)

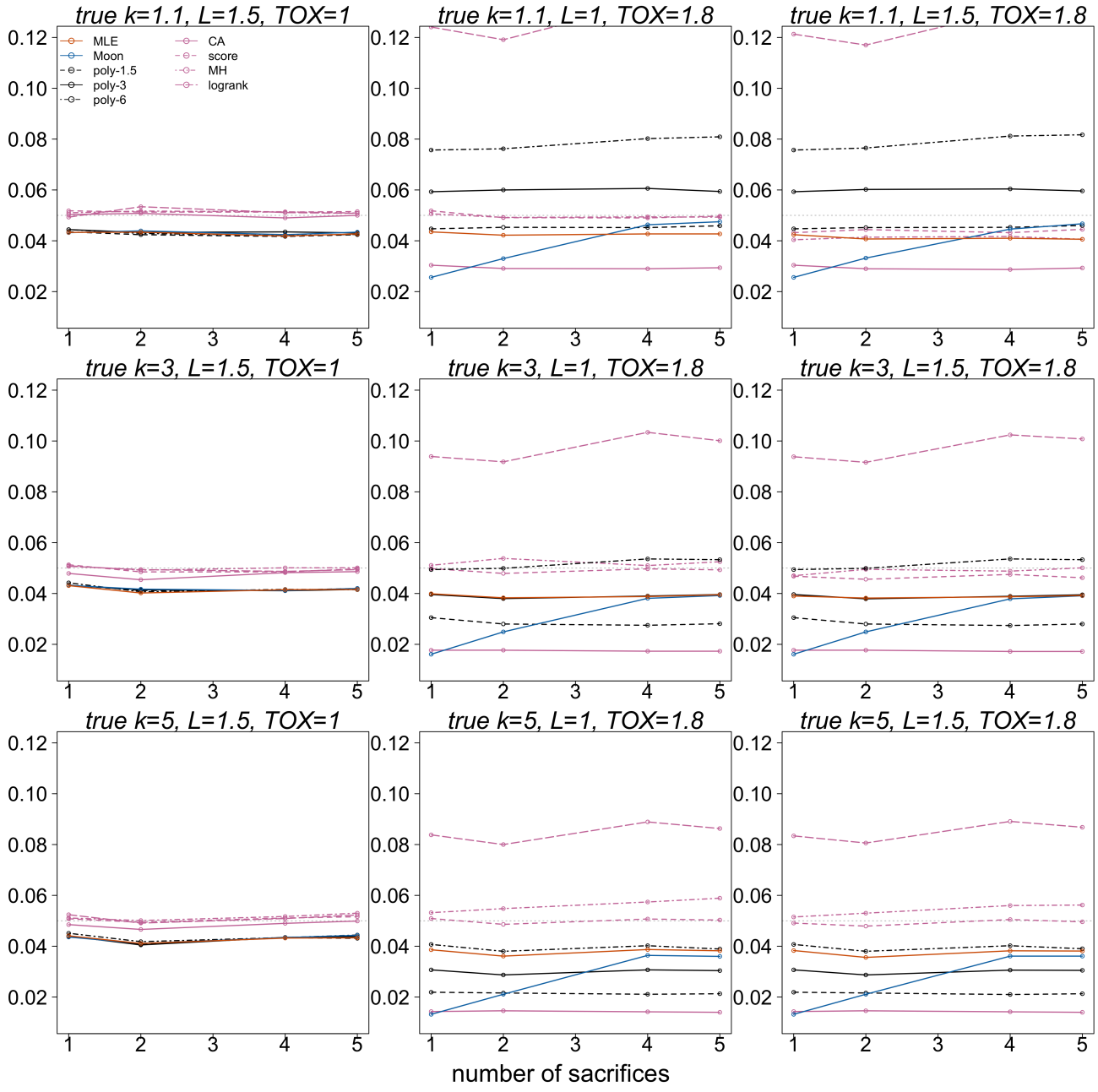


Figure 3.16: Power: $p_0 = 0.15$, varying k_T, lethality, toxicity (log-logistic tumor hazard)

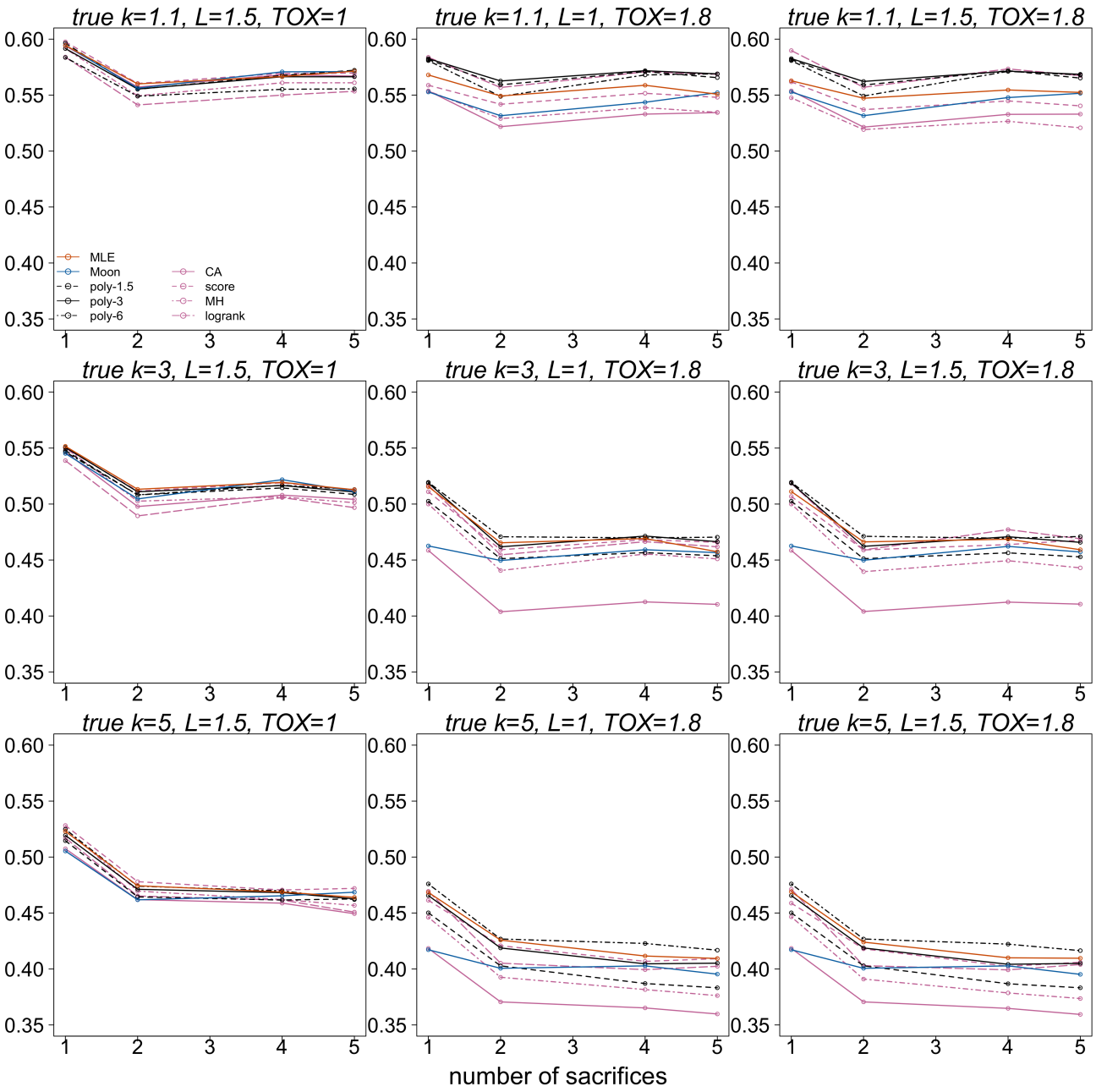


Figure 3.17: Type I error: $p_0 = 0.15$, varying k_T, lethality, toxicity (Gompertz tumor hazard)

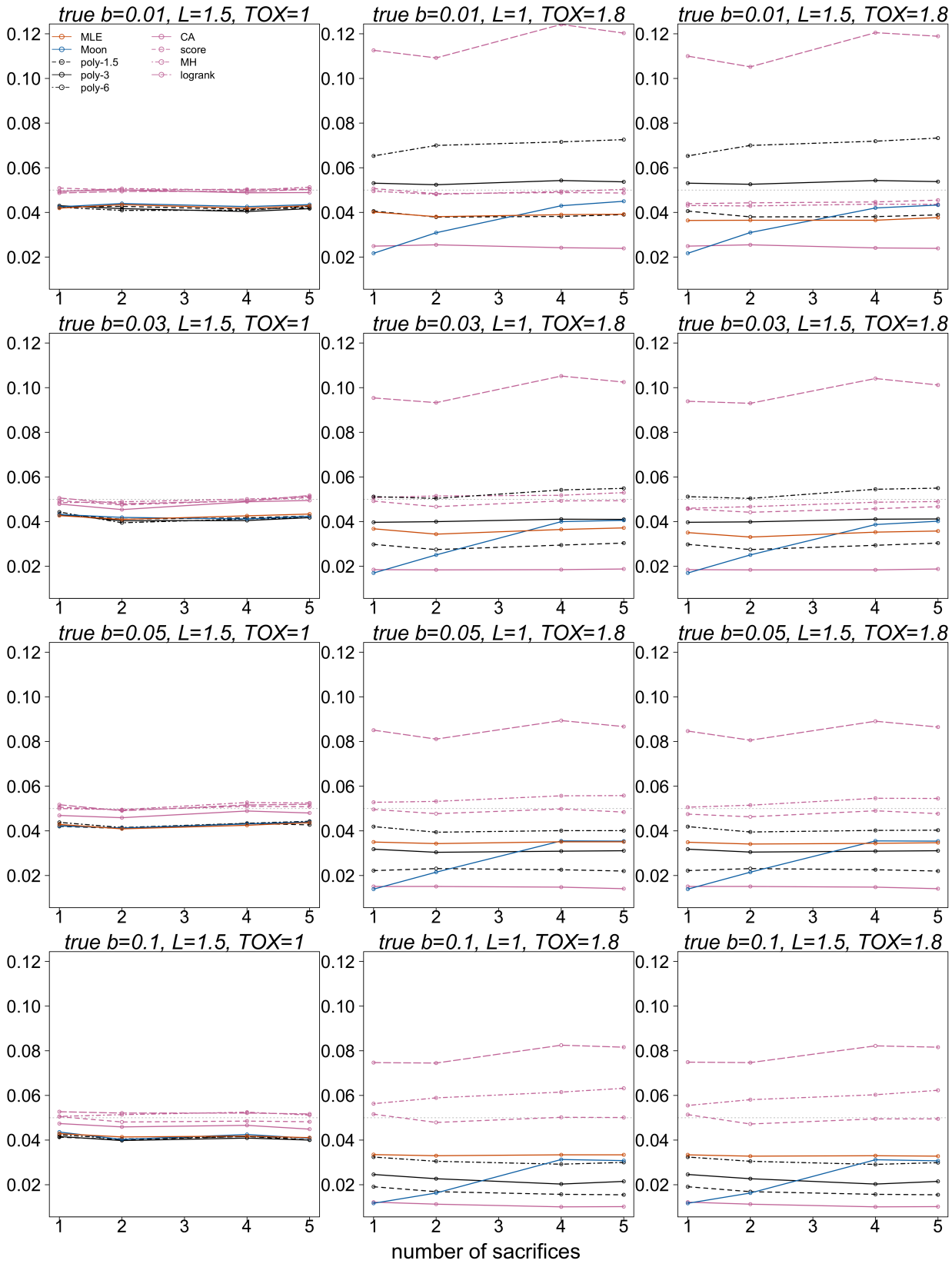
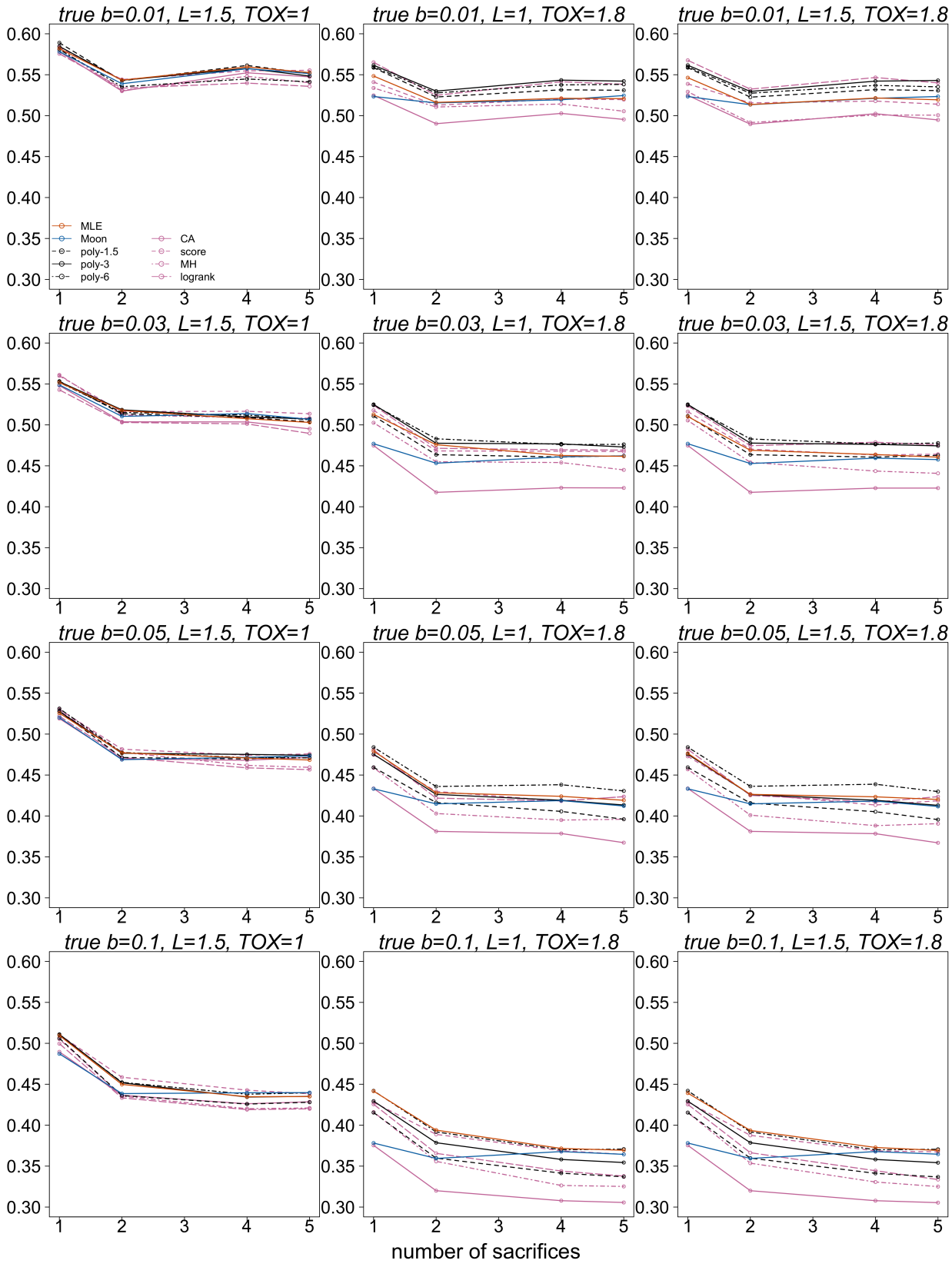


Figure 3.18: Power: $p_0 = 0.15$, varying k_T, lethality, toxicity (Gompertz tumor hazard)



Weibull ($k = 1$), the logrank type I error is greater than 0.12 (Figure 3.11). Under most simulation settings, the score and MH tests maintain 0.05 level better than the poly-k tests, which we have noted previously fail to achieve nominal 0.05 at sample sizes lower than 100. For lower incidence cancers (Figure 3.13), the logrank and MH tests both have inflated type I error across lethality, toxicity, and sacrifice settings, with treatment toxicity inducing the greatest inflation of logrank type I error. This behavior becomes worse for higher true k.

Power (calibrated for tests with nominal 0.05 level) is similar across all tests in the absence of treatment toxicity, for fixed true k and number of sacrifices (left two columns of Figure 3.12). An exception to this is that the CA test generally has lower power compared to the other tests under the same settings when true $k \geq 1.5$. The poly-k test can also have reduced power when there is extreme departure from its modeling assumptions, i.e. the poly-6 has lower power than other tests when true $k = 1$ as does the poly-1.5 when true $k = 6$.

In the presence of treatment toxicity, poly-$\hat{k}_{MLE}$ test power surpasses that of the poly-$\hat{k}_{Moon}$ test, most noticibly at small numbers of sacrifices, but also to a lesser degree up through 4 sacrifices. (See Figure 3.12, right two columns.) This difference between the two tests becomes more pronounced for higher choices of true k. It is worth noting that uncalibrated power differs even more between these two tests: since the poly-$\hat{k}_{Moon}$ test is very conservative when there are few interim sacrifices, calibrating against this lower type I error therefore increases the power. For example, in the presence of both lethality and toxicity, when true $k = 3$ and there is one interim sacrifice, calibrated power is 0.446 for poly-$\hat{k}_{Moon}$ and 0.464 for poly-$\hat{k}_{MLE}$; uncalibrated power is 0.321 for the Moon method and 0.407 for the MLE-based test.

With treatment toxicity, the poly-$\hat{k}_{MLE}$ test has comparable power to the fixed-k poly-k tests, and tends to have power most similar to the poly-k test with correctly-specified k. When there is toxicity, the CA test has low power over all k, lethality, and sacrifice settings; this is consistent with what has been seen previously in the literature. The difference between

the CA test and the other tests' power becomes greater for higher true k. The logistic score test, on the other hand, tends to have slightly lower power compared to the poly-$\hat{k}_{MLE}$ test when true k is low, and compares more favorably when k is higher. The poly-k tests achieve higher power than the logistic score and MH. The logrank test is underpowered (calibrated for 0.05 level) compared to the other tests (Figure 3.12). The only test that tends to have worse power under treatment toxicity is the CA test. In general, all the tests have lower power for higher true k compared to lower k. (All numeric results are available in tabular form in Appendix C.3.)

For lower incidence cancers ($p_0 = 0.05$) when there is treatment toxicity, the MLE-based test still has higher power than the Moon et al. method for low numbers of sacrifices (Figure 3.14). Under some scenarios, and when there are at least 4 total sacrifices, the poly-$\hat{k}_{Moon}$ test has higher power than the poly-$\hat{k}_{MLE}$; e.g. when true $k < 3$. For these rare cancers, the CA test again has low power compared to the poly-k type tests. The logistic score test has similar power to the poly-$\hat{k}_{MLE}$ test, having slightly higher or lower power than the MLE-based test depending on the true value of k. All tests have diminished power in this setting compared with there being higher cancer incidence (the $p_0 = 0.15$ setting discussed above).

For the 2-sided versions of these tests, type I error and power are very similar among tests in the absence of toxicity. When there is treatment toxicity, the 2-sided $poly - \hat{k}_{MLE}$ consistently achieves higher power (adjusted to a 0.05 level) than the 2-sided $poly - \hat{k}_{Moon}$ test. (See appendix Tables C.13 and C.15 for more details.)

Results from simulations with non-Weibull hazards were very similar to those described above. For experiments with fewer than 4 sacrifices, the poly-$\hat{k}_{MLE}$ test continues to maintain level better than the poly-$\hat{k}_{Moon}$ test (Figures 3.15 and 3.17) and has higher calibrated power (Figures 3.16 and 3.18). The poly-$\hat{k}_{MLE}$ is slightly worse at maintaining size when the tumor hazard is Gompertz.

3.3.4 NTP Pulegone example

The NTP Pulegone data set provides a real-world example for application of these methods. Recall that this was a 4-dose 2-year experiment with a single terminal sacrifice. The data were described in detail in sections 2.4 and 2.5.2. We concluded after Chapter 2 that assuming a Weibull family of densities allows a great deal of flexibility, so it is reasonable to focus on estimating the Weibull parameters. In Chapter 3 we have further seen that the poly-$\hat{k}_{MLE}$ performs as well as a poly-k with correctly specified k. Thus it is interesting to see how the poly-$\hat{k}_{MLE}$ test would compare in the context of the results presented in section 2.5.2.

Our previous application of CA-derived hypothesis tests suggested that the observed increase in hepatocellular carcinomas for higher dose groups was statistically significant at $p < 0.05$. However, not all underlying hazard assumptions return this result (Table 2.1). Applied to the liver cancer outcome, the poly-$\hat{k}_{MLE}$ returns $p = 0.0503$, based on the MLE $\hat{k} = 0.613$ in the poly-k test. This $\hat{k}$ is a rather low estimate and was not one of the possibilities for k that we considered earlier. We woud not reject H_0 at the 0.05 level with this borderline p value (which may not be so borderline, depending on how we handle the issue of multiple testing). Here, we are in agreement with the NTP assessment that there is not enough evidence to associate pugelone with liver tumors (although their report did conclude that it is associated with increases in the broader category of "hepatocyte cellular alteration"). [31]

Our earlier hypothesis test results on pugelone and bladder cancer are found in Table 2.2; most tests produced very low p-values. For this outcome, the $poly - \hat{k}_{MLE}$ yields $p = 0.0019$, after estimating $\hat{k} = 0.46$, another low estimate. This result is in keeping with the low end of our earlier poly-k test results, as well as with the $\mathrm{CA}_{\text{log-logistic}} - k$ results. It is also consistent with the NTP conclusion that the association of pugelone with bladder cancer is statistically significant.

3.4 Discussion

We have presented a new testing method for carcinogenicity experiments based on Bailer & Portier's poly-k test, in which we use maximum likelihood to obtain an estimate of the requisite k parameter. The new test performs comparably, as measured by type I error and power, to the poly-k with correctly specified k in a variety of simulated settings. Considering that k is not typically known a priori in practice, this is a very favorable result. In the presence of toxicity, $poly - \hat{k}_{MLE}$ matches or outperforms $poly - \hat{k}_{Moon}$ in most scenarios. In our simulations, when there are at least 4 sacrifices (3 interim), the $poly - \hat{k}_{Moon}$ test is at its best advantage. As mentioned earlier, this is precisely the setting in which Moon et al. present their generalized poly-k test [26]. Even for experiments with 4-5 total sacrifices, the $poly - \hat{k}_{MLE}$ test often has similar size to the $poly - \hat{k}_{Moon}$ test; however, the Moon et al. method does come closer to nominal level when true k is rather small and there are at least 4 sacrifices. Finally, even when there are 4-5 sacrifices, the MLE method usually has equivalent or better calibrated power than the $poly - \hat{k}_{Moon}$ test.

Several aspects of Moon et al.'s approach are improved upon by our method. Their method, perhaps non-intuitively, depends on equating a non-parametric MLE to one based on parametric assumptions, and does not use the full data likelihood. Further, they set the observation time for all tumor-bearing animals to the maximum study time ($F(t_{max})$), which discards information when sacrifices are available (see equation 3.1). This may have been motivated to avoid bias due to differential mortality, and specifically by the way that the α_{ij} formula employed by the poly-k test treats tumor-bearing animals. The form of their LCTIR_1 is attractively simple and incorporates the poly-k time-at-risk adjustment α_{ij}'s in a clear way. Moon et al.'s method can do a reasonable job estimating k, provided that the experimental setting incorporates at least 3 interim sacrifices. However, as our results show, it is not an optimal approach to this estimation problem. The MLE-based approach we have introduced implements a full likelihood and thus is able to utilize more of the information

provided in the data, avoiding a reliance on interim sacrifices.

Our method produces usable estimates even in the absence of serial sacrifice. Assuming tumor onset data that can be modeled using Weibull hazards, obtaining a reasonably unbiased estimate of k makes it possible for poly-$\hat{k}_{MLE}$ to avoid the previously-noted problems that the poly-3 test encounters in the presence of treatment toxicity. It is of interest whether a different serial sacrifice experimental design would result in an improved estimate.

Chapter 4

SERIAL SACRIFICES IN CARCINOGENICITY EXPERIMENTS

4.1 Introduction

The newly introduced poly-$\hat{k}_{MLE}$ test compares favorably with competitor methods in the setting of 2-year animal carcinogenicity experiments. While this test does not require interim sacrifices, it benefits from at least one mid-study look at the tumor onset data. Despite evidence that the MLE of k tends to be less biased when there are more interim sacrifice times (section 3.3.2), this improvement in estimation nevertheless did not clearly translate to the poly-$\hat{k}_{MLE}$ test having improved type I error or power in our earlier simulations (section 3.3.3). Of course, our earlier simulations considered just one serial sacrifice design type.

For serial sacrifices, the simulations in Chapter 3 used a group sample size of $N = 60$, considered between 0 and 4 interim sacrifice times, and held the total animals that would be sacrificed over the course of the experiment fixed at 10-12 per treatment assignment. It is worth investigating whether it is possible to achieve better test performance under different settings of these factors, either by increasing sample size or the number of animals sacrificed per time point. It is possible that for a fully parametric MLE based test, the best use of resources is to devote as many animals as possible to a terminal sacrifice experiment. McKnight & Crowley pointed out that interim sacrifices are required for non-parametric identifiability of tumor incidence rates [23] (and correspondingly, the Moon et al. estimator relies on interim sacrifices).

Interim sacrifices help address concerns with identifiability. Competing risks that apply evenly across treatment groups still allow us to treat death with tumor as a proxy for tumor incidence. When tumors are instantly lethal, tumor prevalence is estimated from deaths with

tumor. However, when tumors are lethal - but not known to be instantly lethal, as in our setting - tumor prevalence is not identifiable [23]. Interim sacrifices provide additional information about tumor onset in the presence of competing risks, although excessive sacrifices can affect our ability to estimate k_T, simply by reducing the available sample at later time points. Ideally, an experimental design should strike a balance between obtaining additional sacrifice information and preserving a healthy proportion of natural deaths.

Dewanji (2005) has noted that "Design issues regarding the optimal sacrifice schedule are of some interest, but not much work has been done on this." [10] In our review of the topic literature, that certainly still seems to be the case.

Bergman and Turnbull (1983) [4] looked into design options, including sequential designs with stopping rules based on the number of tumor / no tumor determinations. They developed the theory of asymptotic efficiency of such designs. However, such sequential designs have not been popular in practice, in part because such methods require rapid evaluation of pathology information during the course of the study [36]. Certainly none of the NTP examples we have examined have incorporated this method.

Borgan, Liestøl, & Ebbesen (1984) compared several experimental designs for investigating time to event data, concluding that serial sacrifices result in efficiency gains over the simple survival experiment [6]. Our data suggest this as well, with a reduction in variability of the MLE $\hat{k}$ as sacrifices are introduced; see for example the last column of Table C.1 as for any choice of true k_T we move from 1 to 2 total sacrifices. When H_0 is true, there is no further reduction in SD, but under H_a there is some reduction of variability with more serial sacrifices, particularly for $k_T \in (3, 6)$, where $h^T(t)$ changes more over time. (See bottom half of Table C.1.)

The sacrifice design Borgan et al. described had equal duration intervals between sacrifices, with animals assigned to a sacrifice time at start of experiment, with larger numbers of animals assigned to later interim sacrifices. The allocation ratio for assigning animals to a

sacrifice time is based on i^2 for interval $i = 1, ..., I$, e.g. with 4 sacrifice times (3 interim sacrifices), allocation of animals is 1:4:9:16 across the four sacrifices. [6] Note that by assigning animals to their sacrifice group a priori, the number that will be interim-sacrificed is not fixed ahead of time but depends on other mortality. The inability to predict ahead of time how many animals will be assigned to interim sacrifice is an obvious problem for study planning, as well as for comparing results across experimental designs. With this pre-assignment approach to sacrifice assignment, differential mortality across treatment groups can also result in less sacrifice information in some (usually high) dose groups relative to others. Due to these factors, we have opted to use the NTP standard for sacrifices, in which a fixed number of animals are randomly selected for sacrifice from the available sample at the end of each interval; for reasonable numbers of sacrifices and assuming that all cause mortality is not especially high, the number of animals chosen for sacrifice is not itself subject to a random process.

Portier (1991) reports results from a small simulation study using goodness-of-fit as an optimality criterion for comparing different choices of sacrifice times and numbers of animals. For a strong dose-response in tumor incidence, Portier concludes that sacrificing 10 animals each at 50 and 75 weeks is optimal. For less extreme dose-response relationships, sacrificing 15 animals at 60 and 85 weeks is preferred. [36] This study is rather limited in that it examined only 4 tumor onset hazards and 6 choices of experimental design. However, the range of these results does agree with the usual choice of 52 and 78 week sacrifice times in the typical NTP experiment. Our simulations, described below, also consider a serial sacrifice design with 10 animals sacrificed at each of 52 and 78 weeks (Table 4.1).

4.2 Simulation Study

In seeking out an optimal design, we consider overall sample size, number sacrificed per interim time, and number of sacrifice times. We simulate 104 week murine carcinogenicity

experiments with 4 dose groups (0 (control), 0.25, 0.5, and 1) and a control group cumulative tumor incidence of $p_0 = 0.15$. Each scenario is run with 5,000 (under H_a) or 10,000 (under H_0) replicated data sets. Different experimental designs vary interim sacrifice times, treatment arm sample size $N \in (50, 60, 70)$, and number of animals selected for sacrifice at each interim time $n \in$ (10-12,15-16,20-21). Table 4.1 reviews the serial sacrifice settings.

Table 4.1: Serial Sacrifice Design: Varying Group Size and Total Interim Sacrificed

# interim sacrifices	Sacrifice times	N (per tx group)	$n^{\ddagger}$ per interim sacrifice	Total sacrificed (interim)
0	104	50 *	—	—
0	104	60	—	—
0	104	70	—	—
1	52, 104	60	10	10
2	52, 78, 104	60	5	10
3	52, 78, 92, 104 †	60	4	12
4	52, 65, 78, 91, 104	60	3	12
1	52, 104	60	15	15
2	52, 78, 104	60	8	16
3	52, 78, 92, 104	60	5	15
4	52, 65, 78, 91, 104	60	4	16
1	52, 104	60	20	20
2	52, 78, 104	60	10	20
3	52, 78, 92, 104	60	7	21
4	52, 65, 78, 91, 104	60	5	20
1	52, 104	70	20	20
2	52, 78, 104	70	10	20
3	52, 78, 92, 104	70	7	21
4	52, 65, 78, 91, 104	70	5	20

* Varying sample size when there is only terminal sacrifice allows us to separate the effect of additional sample size from effect of having interim sacrifices.

† These sacrifice intervals are the NTP standard, according to Dinse [11].

‡ For each group of 4 simulations, n was selected such that the total interim-sacrificed subjects over the whole experiment was similar.

With the exception of this variation in experimental design, data generation details are

identical to what was described in section 3.2. Tumor onset is simulated using Weibull hazards ($k \in (1, 1.5, 3, 6)$; we noted in our results in Chapter 3 that comparable type I and power results for these tests are achieved under Weibull, Gompertz, or log-logistic hazards.

Our focus is on comparing the bias and variability of the two k estimation procedures introduced thus far, as well as assessing the type I error and power of the poly-$\hat{k}$ tests implemented using these estimates. Moon et al.'s estimator was described in section 3.1.2. Our maximum likelihood based estimation of k was described in section 3.1.3.

4.3 Results

4.3.1 Simulation Results: Bias

Figures 4.1 and 4.2 show the mean estimates of k obtained using either maximum likelihood or Moon et al.'s method. We compare experimental designs differing by number of sacrifices (i.e., total number of study intervals), total sample size N per treatment group, and the total number of animals sacrificed per treatment assignment over the duration of the study (shown in the last column of Table 4.1). For these simulations, we compare cases where there was tumor lethality but no treatment toxicity, no tumor lethality but moderate toxicity, and both tumor lethality and toxicity. (Detailed numeric results from these simulations are available in tabular form in Appendix D.1.)

Across all choices of true k parameter, whether H_0 is true or not, we again see the pattern of behavior that was discovered in Chapter 3. Mean $\hat{k}_{Moon}$ is 0 with no interim sacrifices, becoming less biased with an increasing number of sacrifices, and tending to underestimate k until 3-4 sacrifices (2-3 interim). The best scenario for $\hat{k}_{Moon}$ is generally an experiment with 4 sacrifices. By contrast, the maximum likelihood based method tends to overestimate k, but is able to produce much less biased estimates even with few sacrifices. For example, with 2 total sacrifices (1 interim), across all true k, $\hat{k}_{Moon}$ underestimates the parameter by more than $\hat{k}_{MLE}$ overestimates it; taking for instance true $k = 1.5$, when 20 animals are

sacrificed at 52 weeks, the Moon et al. estimate is on average 0.54 lower than the true k, whereas the MLE is 0.25 higher.

At 1 sacrifice, we focus on the MLE method, as the Moon et al. method has very high bias and was originally intended to be used with interim sacrifice data. With no interim sacrifices, we can examine the effect that simply increasing sample size has on $\hat{k}$. There is a measurable reduction in bias of the MLE with greater sample size. For each choice of true k parameter, this reduction is larger when H_0 is true than it is in the corresponding set of simulations when H_a is true. This is partly accounted for by the fact that the bias for $N = 50$ is also lower when H_a is true; this initial bias increases for larger values of true k. This is likely due to a lower rate of tumors early in the study, as we have seen that Weibull with $k = 6$ models later onset tumors.

Introducing an interim sacrifice produces a sizeable reduction in bias for $\hat{k}_{MLE}$. (As we noted in the Chapter 3 results, the exception to this is when true $k = 6$, for which a single sacrifice at 52 weeks is too early, likely taking mostly non tumor-bearing animals out of the experiment.) A sample size of $N = 60$ with 10 interim sacrifices shows lower bias than a sample size of 60 with no interim sacrifices. A study design with $N = 60$ and 20 sacrifices reduces bias even further, despite reducing the overall sample size available for the duration of the experiment. Increasing N to 70, but limiting total interim sacrifices to 20 animals, reduces bias under some settings, but not by a large amount: Under both H_0 and H_a, for most choices of true k, the mean $\hat{k}_{MLE}$'s are very similar between $N = 70$ and $N = 60$ with 20 sacrificed animals. A similar pattern of bias reduction becomes established for $\hat{k}_{Moon}$, particularly for high numbers of sacrifice times. Once at least 3 sacrifice intervals are introduced, a major difference for $\hat{k}_{Moon}$ is that conducting 20 sacifices with $N = 70$ represents a much greater reduction in bias than seen for the MLE.

Although increasing the number of sacrifice times (but sacrificing the same number of animals overall) further reduces the bias of $\hat{k}_{Moon}$, the same is not true for $\hat{k}_{MLE}$, which

remains relatively stable across numbers of sacrifices greater than 1. Interestingly, $k = 6$ is an exception, with the initial interim sacrifice actually increasing the bias of $\hat{k}_{MLE}$. With few early tumors under $h^T(t) = \text{Weibull}(k = 6)$, it appears that introducing a single sacrifice (at 52 weeks) yields little gain in information, and in fact removes 10 animals from the study before they can provide useful tumor onset data. For $\hat{k}_{Moon}$, interim sacrifices make it possible to produce any kind of estimate beyond 0, so we see an improvement associated with the first interim sacrifice even under $k = 6$; however, it is not as sharp as it is for lower values of true k.

Variability of the estimators was assessed by calculating standard deviation and mean squared error (MSE) at each setting (see Figures 4.3 and 4.4, and Tables in Appendix D.1.) The MLE $\hat{k}$ becomes less variable as sacrifices are introduced. Under H_0, increasing the number of interim sacrifices beyond 1 does not reduce variability further. Under H_a introducing more interim sacrifice times does result in a rather small reduction in variability, particularly for high values of true k. For instance, when true $k = 6$ under H_a, with treatment toxicity but no lethality, 60 animals per group and 10 interim-sacrificed, the standard deviation for $\hat{k}_{MLE}$ decreases from 2.439 with 1 interim sacrifice to 2.233 with 4 interim sacrifices. By contrast, the Moon et al. estimate of k tends to be more variable with increased sacrifices, up through 3 interim sacrifices. For $k < 6$, the MLE becomes slightly less variable with larger numbers of animals sacrificed at interim looks, as well as with an increase in overall sample size. This is also true of $\hat{k}_{Moon}$, given at least 2 interim sacrifices.

Of particular interest is how variable the two $\hat{k}$'s are compared to one another. In general, for low numbers of sacrifices, the $\hat{k}_{MLE}$'s are more variable than the $\hat{k}_{Moon}$'s under the same conditions. By 3-4 sacrifice times, the MLE becomes less variable than the Moon et al. estimate. The case where true $k = 3$ under H_0 (treatment toxicity but no lethality) is typical of this pattern: For experiments with 60 animals and 10 interim-sacrificed animals, SD($\hat{k}_{Moon}$)=0.37 and SD($\hat{k}_{MLE}$)=1.92 when there is 1 interim sacrifice; for 2 interim sac-

rifices, SD($\hat{k}_{Moon}$)=1.05 and SD($\hat{k}_{MLE}$)=1.75; and for 3 interim sacrifices, SD($\hat{k}_{Moon}$)=1.95 and SD($\hat{k}_{MLE}$)=1.67. Interestingly, by 4 interim sacrifices, the SD's for these estimates often become very similar. MSE varies quite a bit for $\hat{k}_{MLE}$, and even more for $\hat{k}_{Moon}$, when H_0 is true, depending on the number of sacrifice times, sample size, number of animals sacrificed, and the true value of k. (See Figure 4.3.) When there is only 1 sacrifice (0 interim), both estimators have high MSE under most conditions; this is primarily driven by the higher bias we have seen for 1 sacrifice. A dramatic reduction in MSE when there are interim sacrifices is to be expected for the Moon estimator, which is extremely biased (and not recommended) for terminal sacrifice only experiments. For the MLE, introducing an interim sacrifice can result in a large reduction of MSE, which is not wholly accounted for by improvement in bias. With additional interim sacrifice times, $\hat{k}_{Moon}$ becomes less biased but more variable, with a net effect of increasing MSE with increased numbers of interim sacrifices. By contrast, the MLE shows a small increase in MSE with increasing interim sacrifices. Both estimators have lower MSE for larger sample size, and when there are more animals interim-sacrificed. These patterns of behavior are similar under H_a, though all MSEs are lower under the alternative. (See Figure 4.4.) Also, whereas under H_0, the MLE has somewhat higher MSE than the Moon estimator for 2 and 3 sacrifices, under H_a, the two estimators are more similar (no toxicity) or the MLE has lower MSE (toxicity).

Bias, SD, and MSE results are not much changed when we simulate data with either treatment toxicity, lethality, or both. (Compare columns of Figures 4.1, 4.2, 4.3, and 4.4.)

4.3.2 Simulation Results: Type I Error and Power

Type I error results for the 1-sided poly-$\hat{k}_{MLE}$ and poly-$\hat{k}_{Moon}$ tests in the presence of treatment toxicity are reported in Figure 4.5. The results from the standard poly-k test with fixed $k \in (1.5, 3, 6)$ are also provided for comparison. Given any true value of k, when we consider fixed sample size and vary the number of animals sacrificed, there is no noticeable

Figure 4.1: Mean $\hat{k}$ by sacrifice strategy when H_0 true, varying lethality, toxicity, and k_T (note that n total (interim) sacrifices is 0 when number of sacrifice times = 1)

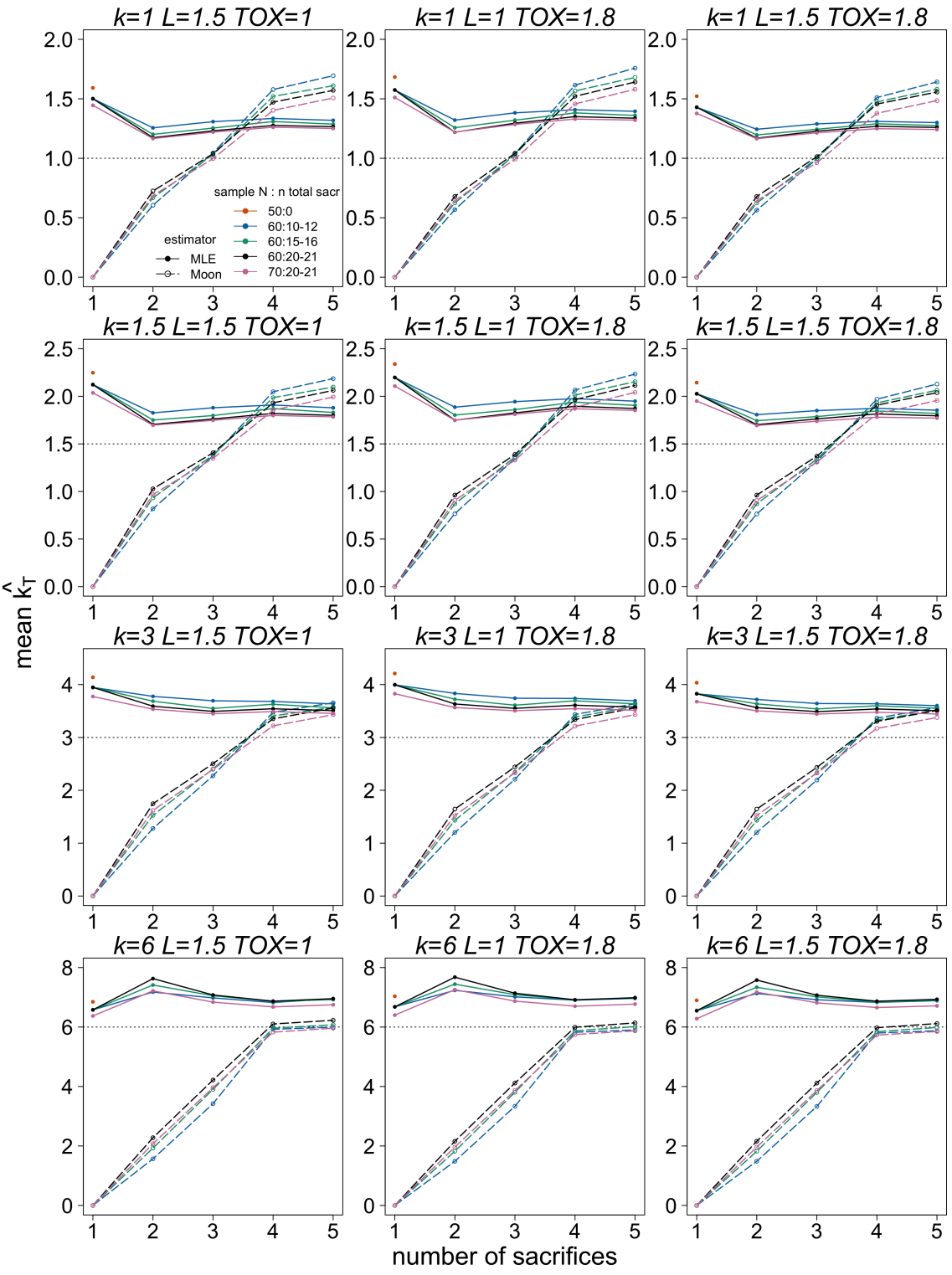


Figure 4.2: Mean $\hat{k}$ by sacrifice strategy when H_a true, varying lethality, toxicity, and k_T (note that n total (interim) sacrifices is 0 when number of sacrifice times = 1)

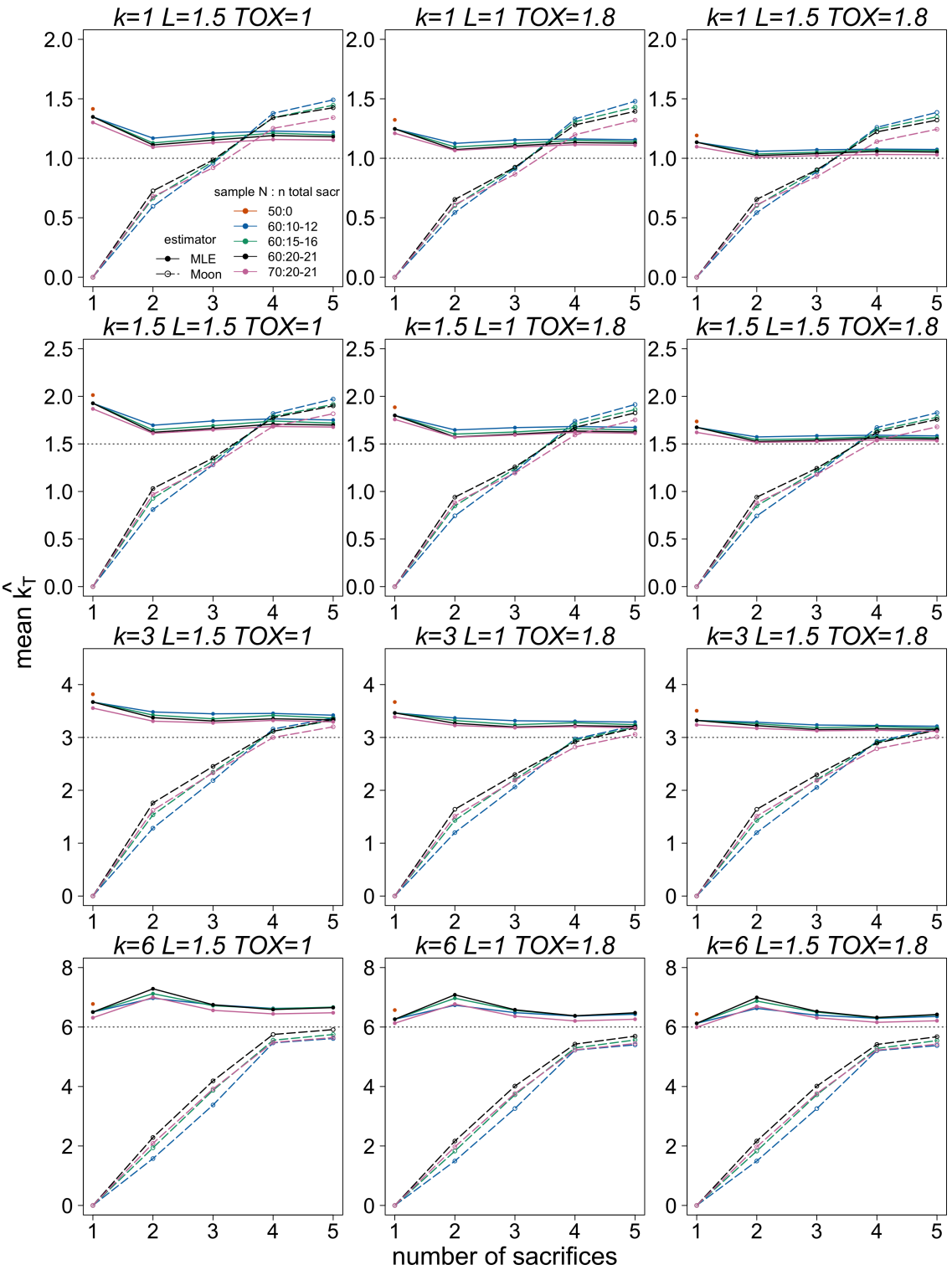


Figure 4.3: MSE of $\hat{k}$ by sacrifice strategy when H_0 true, varying lethality, toxicity, and k_T (note that n total (interim) sacrifices is 0 when number of sacrifice times = 1)

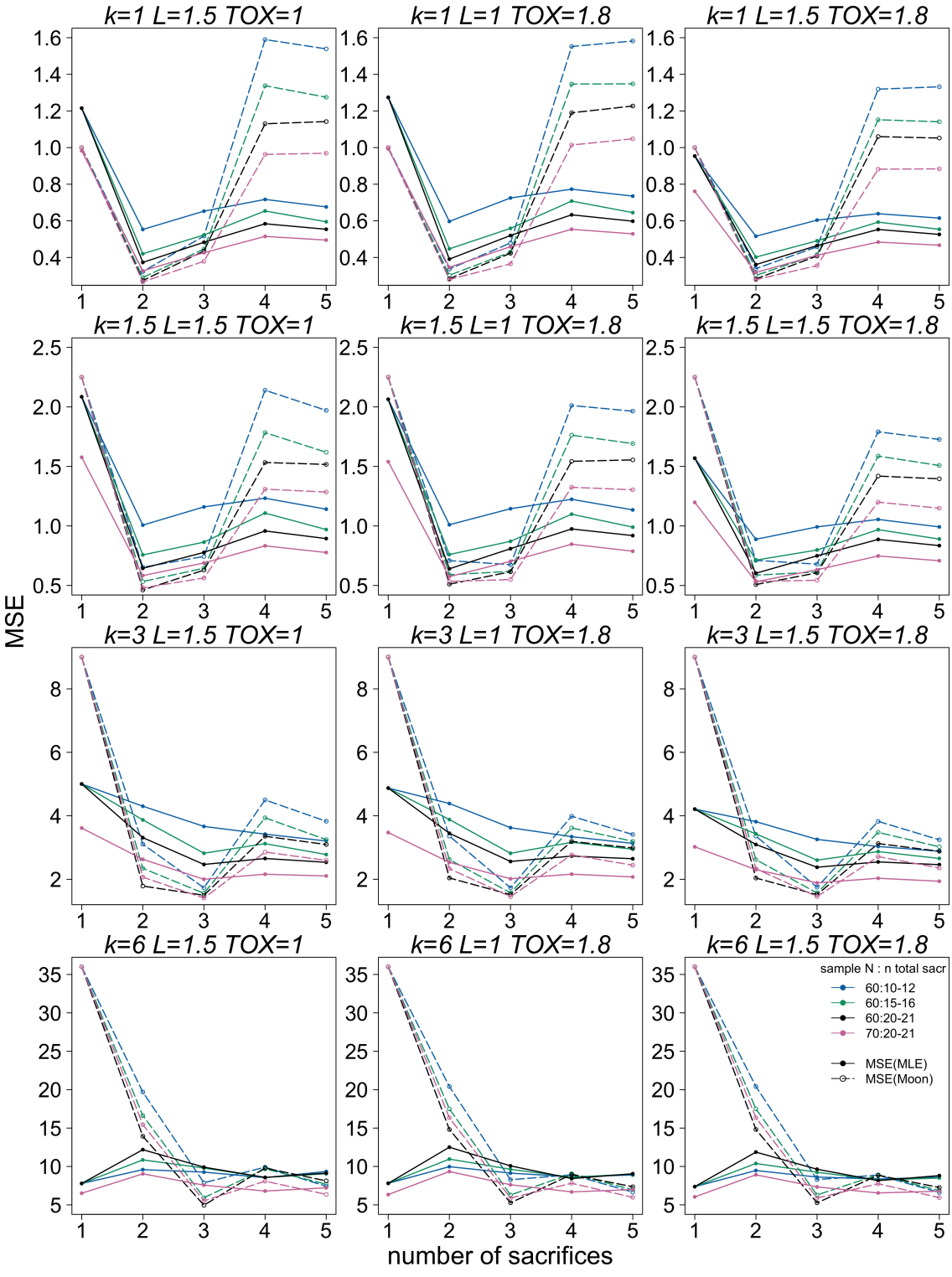


Figure 4.4: MSE of $\hat{k}$ by sacrifice strategy when H_a true, varying lethality, toxicity, and k_T (note that n total (interim) sacrifices is 0 when number of sacrifice times = 1)

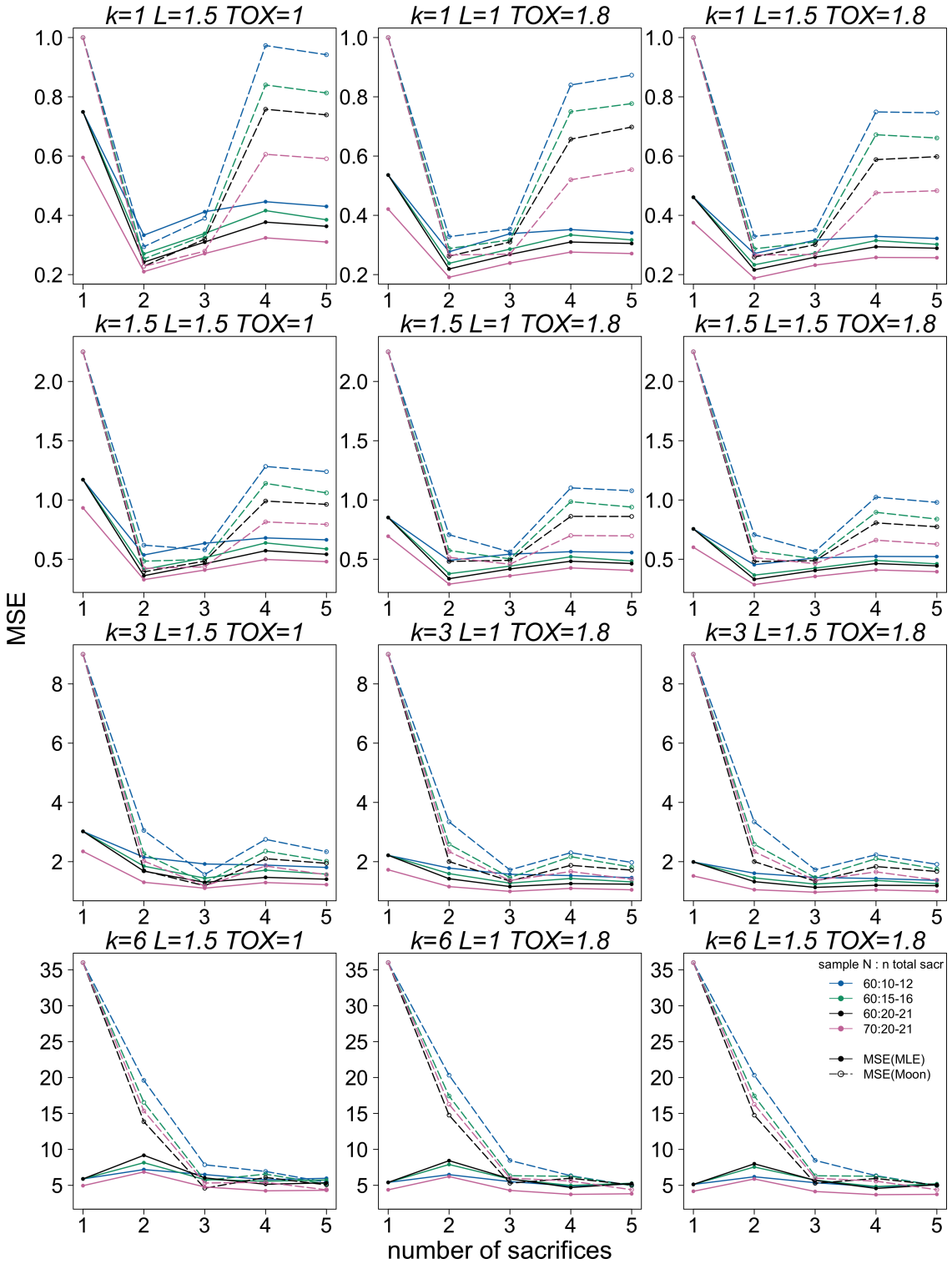


Figure 4.5: Type I error: $p_0 = 0.15$, $L = 1$, $TOX = 1.8$, varying k_T

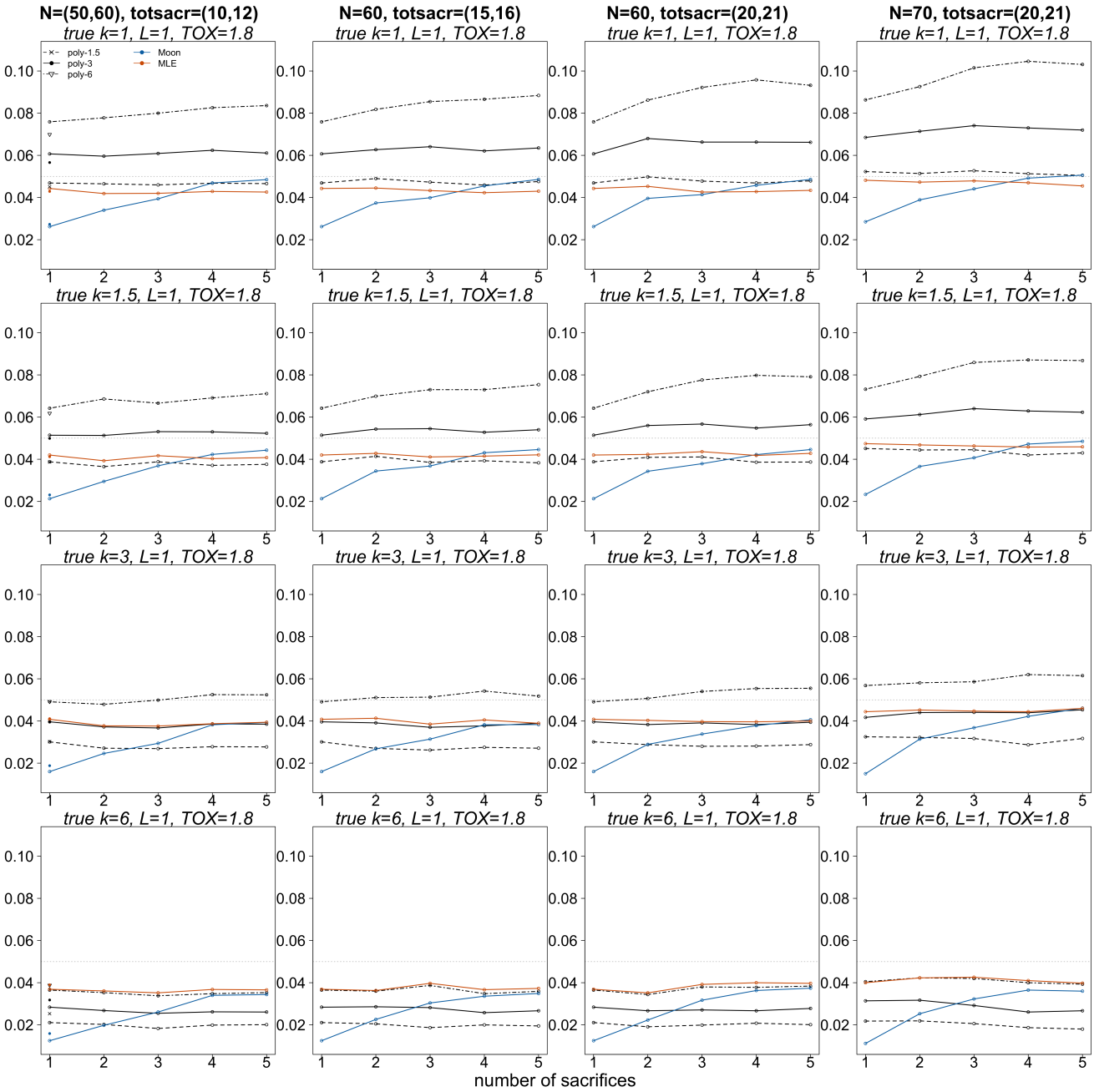


Figure 4.6: Type I error: $p_0 = 0.15$, $L = 1.5$, $TOX = 1$, varying k_T

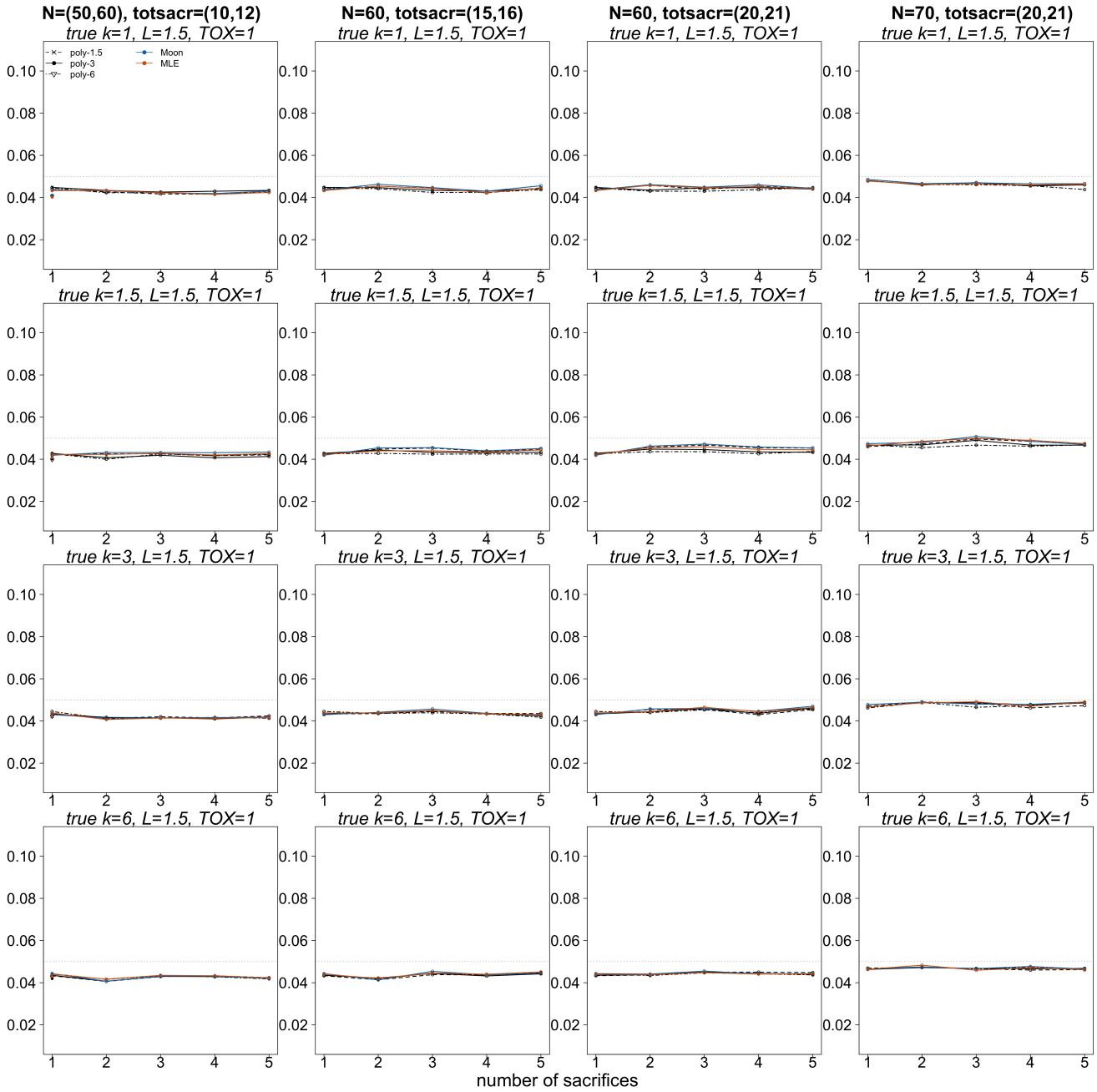


Figure 4.7: Type I error: $p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$, varying k_T

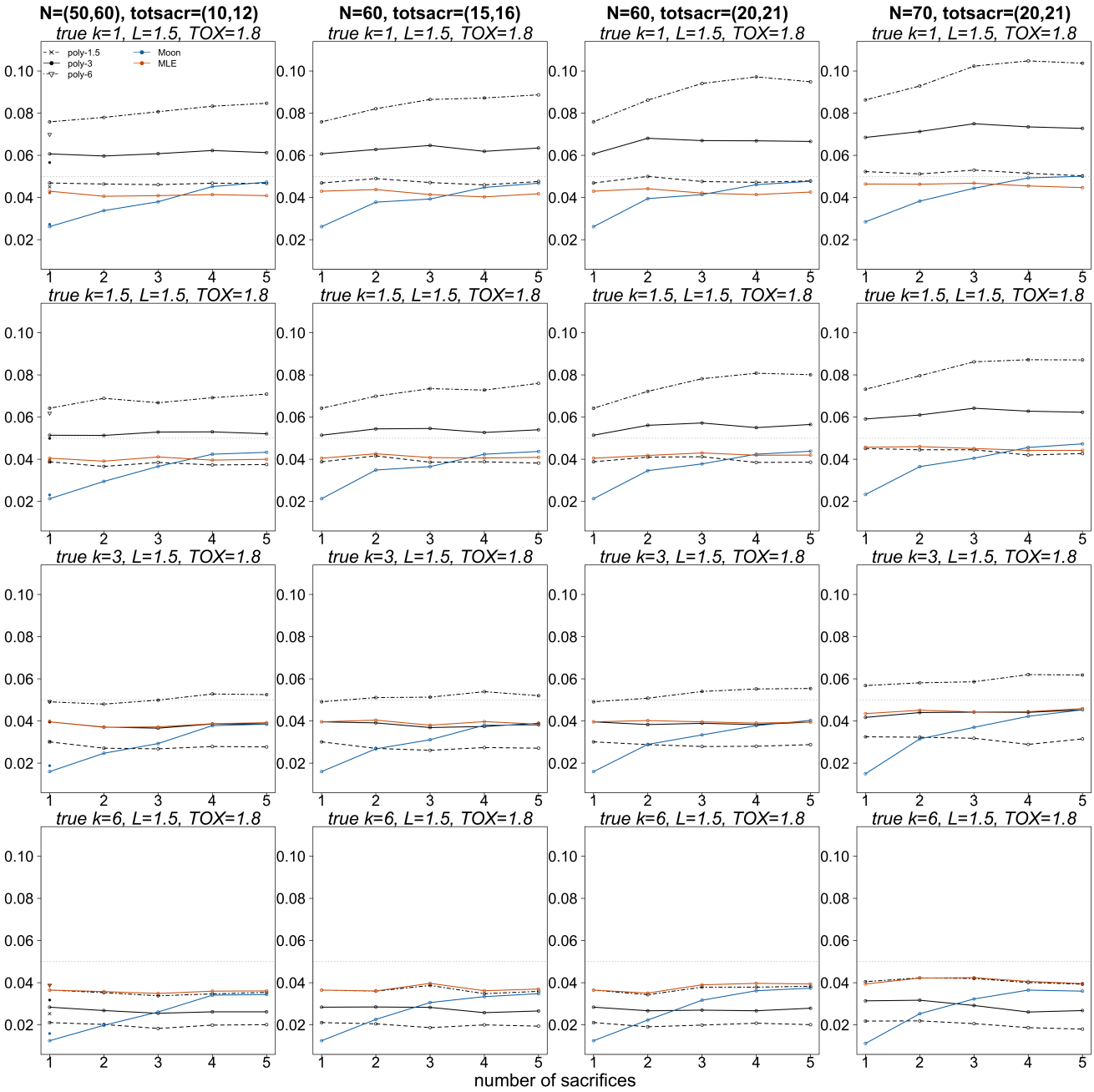


Figure 4.8: Power: $p_0 = 0.15$, $L = 1$, $TOX = 1.8$, varying k_T

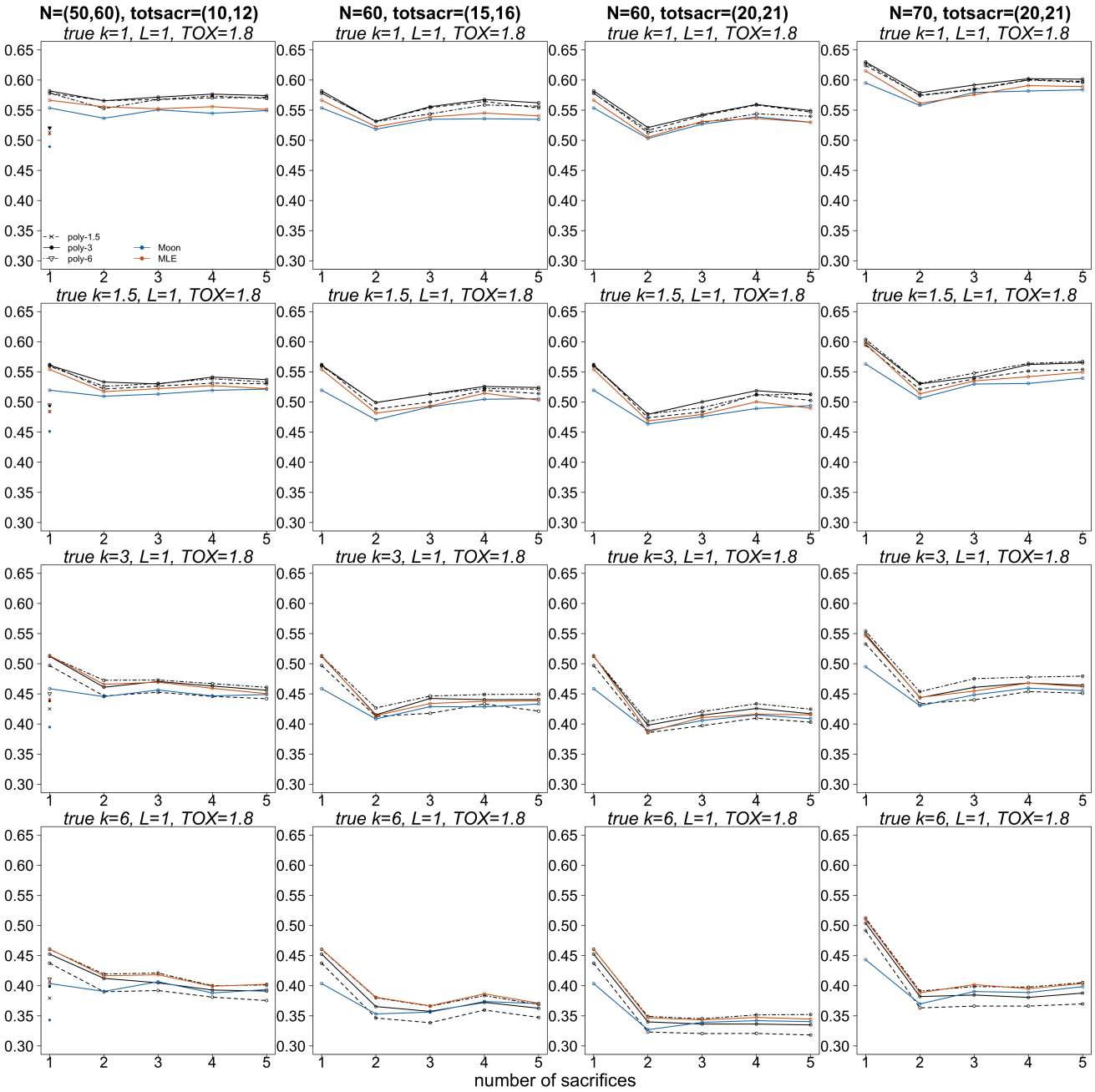


Figure 4.9: Power: $p_0 = 0.15$, $L = 1.5$, $TOX = 1$, varying k_T

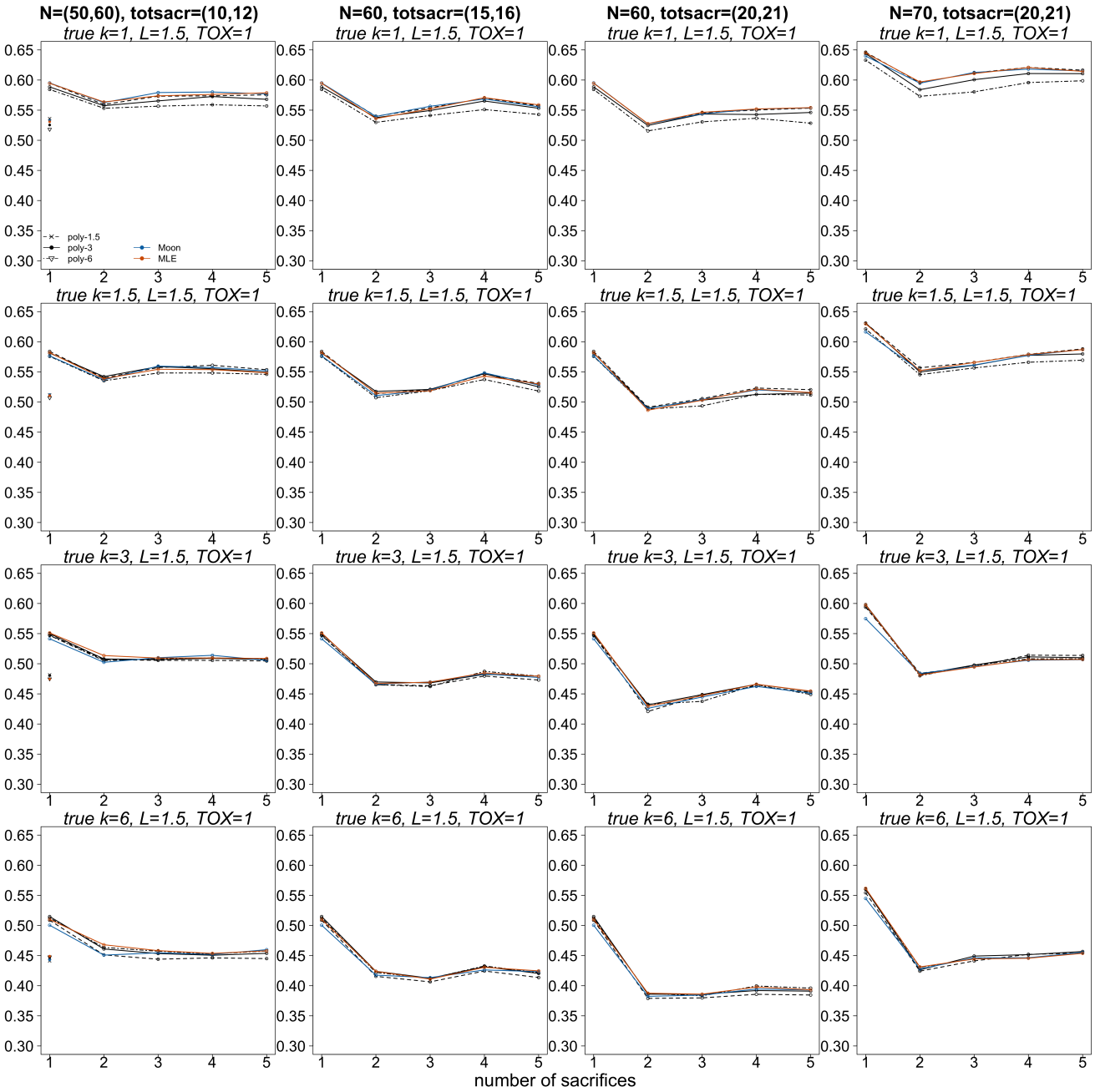
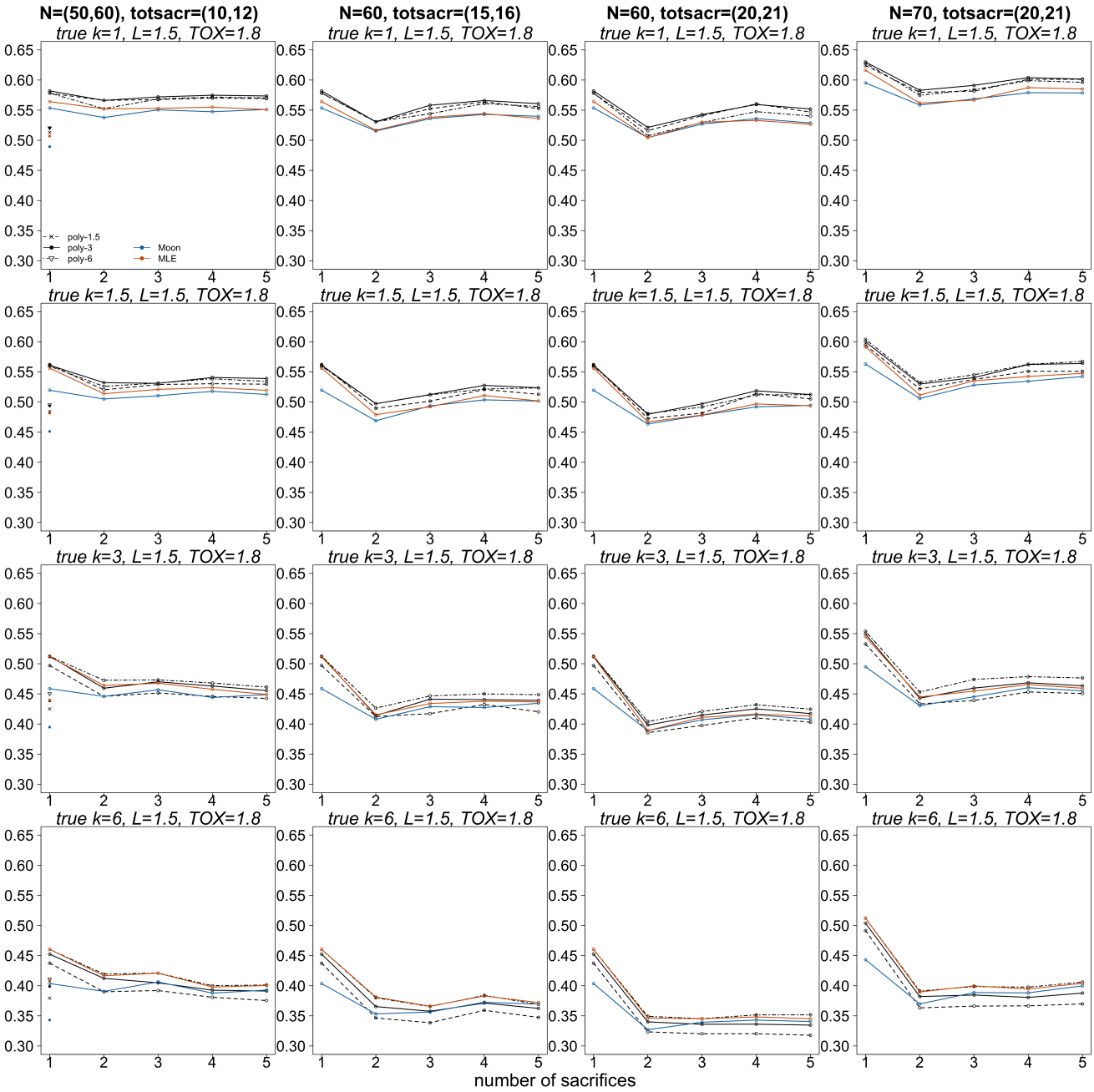


Figure 4.10: Power: $p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$, varying k_T



change in poly-$\hat{k}_{MLE}$ type I error. This is also the case for poly-$\hat{k}_{Moon}$ when it is provided sufficient numbers of sacrifice intervals (or 3-4 total sacrifices). As noted earlier, the poly-k tests fail to reach nominal type I error at sample sizes of 50 and 60. Increasing treatment group sample size to 70 predictably brings tests' size a bit closer to nominal level. The poly-k tests with fixed and correctly specified k show the same type I behavior as the poly-$\hat{k}_{MLE}$, whereas tests which misspecify k as being higher than the truth, inflate type I error and inflate it more for higher numbers of sacrificed animals.

When there is no treatment toxicity (Figure 4.6), the poly-k family of tests is not biased even when k is misspecified. As in Chapter 3 we again see that all fixed and estimated-k tests' type I error results are very similar. When overall sample size is held fixed, there is no notable difference in type I error across differing numbers of sacrificed animals. Increasing the sample size from 60 (with either 10 or 20 total interim-sacrificed animals) to 70 (with 20 sacrifices) does increase the size of the test a bit, but not by much. Presumably the 20 sacrificed animals attenuate the effect of increased sample size.

Power (Figure 4.8) of all the tests considered tends to decrease with increasing numbers of sacrificed animals, and increase when overall sample size is increased. The power of the two tests based on estimating k follows similar patterns to the tests that are not based on estimating k, which suggests that the effect of sample size on power overwhelms any impact that sacrifices have on estimation of k, at least in terms of what may be observed in these simulations. This is still true when the treatment is not toxic (Figure 4.9), and we again see all tests behave much more similarly since the precise specification of k is not important in the absence of toxicity.

4.4 Discussion

Tests' performance characteristics (type I error and power) did not provide any evidence for adopting a particular experimental design over others, aside from increasing overall sample

size. However, examining the bias and variability in the parameter estimation does allow us to differentiate among the options.

In the settings we examined, a single interim sacrifice of 20 animals with a sample size of $N = 60$ per treatment group yields the least biased maximum likelihood estimate of k_T. Under some conditions, a sample size of $N = 70$ improves the estimate, but perhaps not to a degree that would justify the expense of 40 additional animals across a 4-dose study. That said, considering the results reported by Moon et al. (Table III data generated with varying numbers of sacrifices), it is very likely that an experiment with a sample size of 70 animals per treatment group and a total of 30 or 40 interim-sacrificed animals would result in worthwhile improvement. Further, except when true $k = 6$, MSE is reduced when there is at least one interim sacrifice. In the presence of both lethality and toxicity, the poly-$\hat{k}_{MLE}$ was closer to achieving nominal 0.05 level with 1 interim sacrifice compared to none. However, in general, test performance (and in particular power) did not follow the trends in bias or MSE.

If the tumor type of interest is known to be late onset, we saw in Figures 4.1 and 4.2 (bottom row in each) that interim sacrifices may not provide enough helpful information to offset the loss of sample size, at least for the settings we considered. MSE is also not notably improved by interim sacrifices in this example (Figures 4.3 and 4.4 bottom row of plots). In this case, a sample size of $N = 60$ in a terminal sacrifice experiment may be preferable to implementing a serial sacrifice strategy.

Our simulations examined lethality as a constant parameter across dose. It is possible that a difference in test performance (type I or power) would be observed if lethality differed by dose.

Interim sacrifices do provide additional information about time to tumor, helping to bypass non-identifiability problems when death cannot be considered a proxy for tumor onset time. Parametric assumptions are another way to evade the identifiability issues inherent to

this setting, although this can potentially create problems when the model is misspecified. However, by using the data driven estimation approach of the poly-$\hat{k}_{MLE}$ with the rather flexible Weibull parametric family assumption, it may be that there is not much further to gain by adding multiple interim sacrifices to the experimental design. This question is by no means settled, as other simulation settings remain for examination, including timing of sacrifices and experimental designs with a larger proportion of animals interim-sacrificed.

Serial sacrifice experiments are relatively rare. Dewanji (2005) notes that there are few examples in the literature [10]. In the large NTP database we used to obtain the Pulegone example (see section 2.4), there were 574 data sets from experiments of the sort considered here (approximately 2 year toxicology experiments with 3 or 4 dose groups of 50-60 animals). Among those, we found that 303 were terminal sacrifice experiments, 197 had 1 interim sacrifice, 45 had 2 interim sacrifices, and only 29 had more than 2 interim sacrifices. Studies using more than 1 interim sacrifice comprise less than 13% of this sample. Considering that the Moon et al. procedure requires at least 2, preferably 3 interim sacrifices, this suggests a significant challenge to adopting their method for currently available NTP data. The MLE method is less hampered by this problem.

Design complexity and cost are certaintly factors that inhibit use of serial sacrifice designs. Motivating experimentalists to adopt alternative sacrifice designs also requires compelling evidence to convince agency decision makers to make this change. In our selection of simulations, the MLE method does well with 0 or 1 interim sacrifices, making it more likely that the poly-$\hat{k}_{MLE}$ would be an acceptable option for experimenters in this application area.

Chapter 5
CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

The poly-k test introduced by Bailer & Portier (1988) [3], which heavily relies on the Weibull assumption, is among the most common methods used for analyzing 2-year carcinogenicity experiments. Acknowledging that the poly-k and competitor tests have been previously studied only using Weibull hazards, our simulations examined these tests under log-logistic and Gompertz hazards. Simulating under less-extreme non-Weibull hazards did not produce significantly worse performance than using similar Weibull hazards (with $0.5 \leq k \leq 5$). Our results suggest that the Weibull family of hazards, using k parameters that are consistent with what is typically found for real data, is sufficiently flexible to justify using it as a default for data generation in simulation studies. Nevertheless, it is still prudent, when assessing new tests with a strong Weibull assumption, to simulate data from non-Weibull hazards as well, as some tests may be more sensitive to small differences between hazard shapes.

More extreme non-Weibull hazards can severely affect poly-k test performance, but so can similarly shaped extreme Weibull hazards, and furthermore these are not likely to reflect what happens in real experiments. For instance, an analysis by Portier et al. (1986) of historical control data from the NTP suggested that Weibull tumor hazards with shape parameter $1 \leq k \leq 5$ was consistent with real data situations [35]. Considering this, it is notable that our simulations found that the poly-3 test's type I error was not robust to the effects of treatment toxicity when k was misspecified within this range. We introduced two new competitor tests, based on the Cochran-Armitage test like the poly-k, but using either log-logistic or Gompertz hazard assumptions. These tests have similar performance char-

acteristics to the poly-k, and can perform better under some circumstances, but ultimately all of the Cochran-Armitage based tests are similarly affected by departures from their own modeling assumptions. Development of methods to relax these modeling assumptions is well-motivated.

Under a Weibull distribution assumption, we derived the maximum likelihood estimator for the k parameter of the poly-k test, yielding a new poly-$\hat{k}_{MLE}$ test. In simulation studies varying tumor lethality, treatment toxicity, and the true value of the tumor hazard shape parameter k, the new test's type I error and power are similar to the poly-k test with k correctly specified. In the presence of treatment toxicity, the poly-$\hat{k}_{MLE}$ test is superior to poly-k tests with misspecified fixed k, as well as to the original Cochran-Armitage test. As with all poly-k tests, the performance of the poly-$\hat{k}_{MLE}$ does not depend on tumor lethality, unlike the logistic score and logrank tests.

Compared to the k estimation method proposed by Moon et al. [26], $\hat{k}_{MLE}$ is less biased for experiments with low numbers of sacrifices (< 3 interim sacrifices). The poly-$\hat{k}_{MLE}$ test achieves similar or better power than poly-$\hat{k}_{Moon}$ in most simulation scenarios, and under a variety of serial sacrifice experimental designs. Moon et al. suggested that the poly-$\hat{k}_{Moon}$ test only be used when interim sacrifice data are available, recommending that different methods be used otherwise [26]. By contrast, the performance characteristics of the poly-$\hat{k}_{MLE}$ test suggest that no such accommodation is necessary.

Lack of reliance on serial sacrifice information makes the MLE-based test more flexible than the poly-$\hat{k}_{Moon}$ test, allowing it to be adopted across a range of experimental designs, with comparable or better performance than other tests in the poly-k family. We saw this to be true even when tumor onset was modeled using non-Weibull underlying hazards. Whether some serial sacrifice designs may nevertheless be better for poly-$\hat{k}_{MLE}$ performance is an open question. The experimental designs we considered for use with the poly-$\hat{k}$ tests did not show a marked improvement of one serial sacrifice strategy over another. Although the estimates

of $\hat{k}$ did become less biased with larger numbers of sacrificed animals (as well as with larger sample size), this did not translate to improved type I error or power of the poly-$\hat{k}_{MLE}$ test. It may be that larger numbers of sacrificed animals per sacrifice time are necessary to produce a noticible improvement in test performance characteristics. Further exploration of this is certainly warranted.

We also did not consider timing of sacrifices; depending on the actual distribution of tumor onsets, sacrificing animals too early results in culling primarily animals that have not had time to develop tumors yet, whereas sacrificing too late risks capturing only animals that have already developed tumors; an opimal design would sample a mix of tumor-bearing and tumor-free animals. Our exploration defaulted to an NTP standard, but there does not seem to be a consensus in the literature that the question of sacrifice spacing has been fully explored [10].

The poly-$\hat{k}_{MLE}$ test represents a potentially useful contribution to animal carcinogenicity research methodology. The poly-k test is extremely popular for analyses of 2-year experiments, but its good performance relies on the assumption of Weibull hazards with a particular choice of shape parameter k. By relaxing the requirement to pre-specify a choice of k, the poly-$\hat{k}_{MLE}$ test avoids the conservative or anti-conservative behavior of the poly-k test (depending on the direction of model misspecification) that results from mis-specified k in the presence of treatment toxicity. When there is no treatment toxicity, all poly-k tests behave similarly whether they use a pre-specified k or an estimate.

Alternate applications for this methodology include any analysis that involves hidden time-to-failure data. One non-biological category of applications is machine component testing, in manufacturing scenarios where a test is necessarily destructive. Other examples suggested by Bergman & Turnbull include failure testing of photographic film or fire extinguishers [4].

5.2 Future Work

A variety of hazard shapes have been considered in the simulations presented here. Real-world examples, such as the pulegone data set (see section 3.3.4) suggest that Weibull tumor onset hazards with $0 < k < 1$ should also be explored in more detail.

While our new MLE-based estimation of k for the poly-k test performs well compared to existing methods, there are several ways that the approach might be expanded, potentially improving test performance. Our initial implementation of the optimization routine has set the treatment effect parameter η to 0, assuming H_0 to be true. This assumption might be justified by considering that the hypothesis test is comparing the poly-$\hat{k}_{MLE}$ test statistic against its null distribution. However, the estimation problem regarding what k to use with the test does not itself have to be reliant on assuming the null hypothesis. Maximizing the log likelihood over all 7 parameters, allowing η to vary, may yield better estimates of the tumor hazard shape parameter k. Separately, it may also be that fixing the nuisance parameters that presented a particular estimation problem would improve this algorithm. For instance, examining the likelihood surface along the lethality parameter L_{tum} indicates ridge-like behavior, and maximizing the likelihood for several possible fixed settings of L_{tum} is a promising alternative.

Another, possibly more efficient, approach to the standard maximization we have conducted so far is to assume that the unobserved data (tumor onset times) are known and to employ an expectation-maximization (EM) algorithm to obtain the MLE [9].

The MLE-based method may also be extended by incorporating cause-of-death information into the likelihood function. Cause-of-death assessments, during which a pathologist determines whether death was tumor-related, have been made to allow for other analysis methods, including the well-known Peto test [33, 34], which pools the prevalence method (for incidental tumors) and death-rate method (for fatal tumors). Possible drawbacks to including such data in the likelihood include increased cost for the additional pathologic

determination and concerns with the accuracy of pathology assessments. The NTP would also be unlikely, under the current data collection standards, to adopt such an approach.

One might also consider modifying the poly-$\hat{k}_{MLE}$ into an adaptive test, estimating k during the study, after each natural death or sacrifice time, with rules in place for stopping serial sacrifices after sufficient information (according to some set of criteria) is available to obtain an unbiased parameter estimate. Previous literature touching on this topic includes Bergman & Turnbull's more general methods development for stopping rules in serial sacrifice experiments [4].

An alternative to using the poly-k framework is simply to obtain the parametric MLE's under H_0 ($\eta = 0$, maximizing over the other 6 parameters) and H_a (maximizing over all 7 parameters) for use in a likelihood ratio test (LRT). Variations on this method include making different parametric choices for the hazard functions used in the data likelihood. For instance, one could expand the available choices of hazard and consider a super-parametric family of hazards. One ad hoc option for computing a LRT within a larger family would be to separately maximize several possible likelihoods, each relying on hazard assumptions from a defined set, and then to take the maximum from among those maxima. Since the parametric LRT uses the likelihood directly in the test statistic and can be affected by tumor lethality, it is possible that this method would more clearly benefit from an increase in serial sacrifices than we were able to detect for the poly-$\hat{k}_{MLE}$ approach.

A further option for conducting a likelihood ratio test would be to implement the likelihood using a piecewise linear underlying tumor hazard. Or for more flexibility, cubic splines could be used, with a constraint to be twice-differentiable. While these choices would be less reliant on parametric assumptions than, for example, a Weibull hazard, these approaches pose the challenge that such piecewise hazards introduce additional parameters to maximize over. This could result in computational difficulties, and more data would likely also be required to support estimation of more parameters.

BIBLIOGRAPHY

[1] Ahn, H., and R.L. Kodell. "Estimation and Testing of Tumor Incidence Rates in Experiments Lacking Cause-of-Death Data." *Biometrical Journal* 37 (1995): 745-63.

[2] Armitage, P. "Tests for Linear Trends in Proportions and Frequencies." *Biometrics* 11, no. 3 (1955): 375-86.

[3] Bailer, A.J., and C.J. Portier. "Effects of Treatment-Induced Mortality and Tumor-Induced Mortality on Tests for Carcinogenicity in Small Samples." *Biometrics* 44, no. 2 (1988): 417-431, doi:10.2307/2531856.

[4] Bergman, S.W., and B.W. Turnbull. "Efficient Sequential Designs for Destructive Life Testing with Application to Animal Serial Sacrifice Experiments." *Biometrika* 70, no. 2 (1983): 305-314, doi:10.2307/2335545.

[5] Bieler, G.S., and R.L. Williams. "Ratio Estimates, the Delta Method, and Quantal Response Tests for Increased Carcinogenicity." *Biometrics* 49, no. 3 (1993): 793-801, doi:10.2307/2532200.

[6] Borgan, O., K. Liestøl, and P. Ebbesen. "Efficiencies of Experimental Designs for an Illness-Death Model." *Biometrics* 40, no. 3 (1984): 627-638, doi:10.2307/2530906.

[7] Broyden, C.G. "The Convergence of a Class of Double-Rank Minimization Algorithms 1. General Considerations." *IMA Journal of Applied Mathematics* 6, no. 1 (1970): 76-90, doi:10.1093/imamat/6.1.76.

[8] Broyden, C.G. "The Convergence of a Class of Double-Rank Minimization Algorithms 2. The New Algorithm." *IMA Journal of Applied Mathematics* 6, no. 3 (1970): 222-31, doi:10.1093/imamat/6.3.222.

[9] Dempster, A.P., N.M. Laird, and D.B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society Series B (Methodological)* 39, no. 1 (1977): 1-38.

[10] Dewanji, A. "Serial-Sacrifice Experiments." *Encyclopedia of Biostatistics* 7 (2005): 4910-4915.

[11] Dinse, G.E. "A Comparison of Tumour Incidence Analyses Applicable in Single-Sacrifice Animal Experiments." *Statistics in Medicine* 13, no. 5-7 (1994): 689-708, doi:10.1002/sim.4780130530.

[12] Dinse, G.E, and J.K. Haseman. "Logistic Regression Analysis of Incidental-Tumor Data from Animal Carcinogenicity Experiments." *Fundamental and Applied Toxicology* 6, no. 1 (1986): 44-52, doi:10.1016/0272-0590(86)90262-9.

[13] Dinse, G.E. "Testing for a Trend in Tumor Prevalence Rates: I. Nonlethal Tumors." *Biometrics* 41, no. 3 (1985): 751-70, doi:10.2307/2531296.

[14] FDA. "Guidance for Industry: Statistical Aspects of the Design, Analysis, and Interpretation of Chronic Rodent Carcinogenicity Studies of Pharmaceuticals. Draft Guidance." Silver Spring, MD: FDA/CDER, 2001.

[15] Fletcher, R. "A New Approach to Variable Metric Algorithms." *The Computer Journal* 13, no. 3 (1970): 317-22, doi:10.1093/comjnl/13.3.317.

[16] Goldfarb, D. "A Family of Variable-Metric Methods Derived by Variational Means." *Mathematics of Computation* 24, no. 109 (1970): 23-26, doi:10.2307/2004873.

[17] Haseman, J.K. "Statistical Issues in the Design, Analysis and Interpretation of Animal Carcinogenicity Studies." *Environmental Health Perspectives* 58 (1984): 385-92.

[18] Kodell, R.L. "Should We Assess Tumorigenicity With the Peto or poly-k Test?" *Statistics in Biopharmaceutical Research* 4, no. 2 (2012): 118-24, doi:10.1198/sbr.2010.10030.

[19] Kodell, R.L., and H. Ahn. "An Age-Adjusted Trend Test for the Tumor Incidence Rate for Multiple- Sacrifice Experiments." *Biometrics* 53, no. 4 (1997): 1467-74, doi:10.2307/2533512.

[20] Kodell, R.L., and H.S. Ahn. "Nonparametric Trend Test for the Cumulative Tumor Incidence Rate." *Communications in Statistics - Theory and Methods* 25, no. 8 (1996): 1677-92.

[21] Kodell, R.L., J.J. Chen, and G.E. Moore. "Comparing Distributions of Time to Onset of Disease in Animal Tumorigenicity Experiments." *Communications in Statistics - Theory and Methods* 23, no. 4 (1994): 959-80, doi:10.1080/03610929408831298.

[22] Lawless, J.F. *Statistical Models and Methods for Lifetime Data.* John Wiley & Sons, Inc., 2003.

[23] McKnight, B., and J. Crowley. "Tests for Differences in Tumor Incidence Based on Animal Carcinogenesis Experiments." *Journal of the American Statistical Association* 79, no. 387 (1984): 639-648, doi:10.2307/2288411.

[24] McKnight, B. "A Guide to the Statistical Analysis of Long-Term Carcinogenicity Assays." *Fundamental and Applied Toxicology* 10, no. 2 (1988): 355-364, doi:10.1016/0272-0590(88)90321-1.

[25] Moon, H, H. Ahn, and R.L. Kodell. "An Age-Adjusted Bootstrap-Based poly-k Test." *Statistics in Medicine* 24, no. 8 (2005): 1233-44, doi:10.1002/sim.1967.

[26] Moon, H., H. Ahn, R.L. Kodell, and J.J. Lee. "Estimation of K for the poly-k Test with Application to Animal Carcinogenicity Studies." *Statistics in Medicine* 22, no. 16 (2003): 2619-36, doi:10.1002/sim.1444.

[27] Nelder, J.A., and R. Mead. "A Simplex Method for Function Minimization." *The Computer Journal* 7, no. 4 (1965): 308âĂŞ13, doi:10.1093/comjnl/7.4.308.

[28] "2-Year Study Protocol - NTP". Accessed June 19, 2016. http://ntp.niehs.nih.gov/testing/types/cartox/protocols/2year/index.html.

[29] "Download NTP Study Data - NTP." Accessed August 24, 2015. https://ntp.niehs.nih.gov/results/dbsearch/download/index.html.

[30] "Expanded Overview - NTP." Accessed June 19, 2016. https://ntp.niehs.nih.gov/testing/types/stats/expanded/index.html.

[31] "Abstract for TR-563 - Pulegone (CASRN 89-82-7) - NTP." Accessed May 12, 2016. http://ntp.niehs.nih.gov/results/pubs/longterm/reports/longterm/tr500580 /listedreports/tr563/index.html.

[32] Peddada, S.D., and G.E. Kissling. "A Survival-Adjusted Quantal-Response Test for Analysis of Tumor Incidence Rates in Animal Carcinogenicity Studies." *Environmental Health Perspectives* 114, no. 4 (2006): 537-41, doi:10.1289/ehp.8590.

[33] Peto, R. "Editorial: Guidelines on the Analysis of Tumour Rates and Death Rates in Experimental Animals." *British Journal of Cancer* 29, no. 2 (1974): 101âĂŞ5.

[34] Peto, R., M.C. Pike, N.E. Day, R.G. Gray, P.N. Lee, S. Parish, J. Peto, S. Richards, and J. Wahrendorf. "Guidelines for Simple, Sensitive Significance Tests for Carcinogenic Effects in Long-Term Animal Experiments." *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans* no. 2 Suppl (1980): 311âĂŞ426.

[35] Portier, C.J., J.C. Hedges, and D.G. Hoel. "Age-Specific Models of Mortality and Tumor Onset for Historical Control Animals in the National Toxicology ProgramâĂŹs Carcinogenicity Experiments." *Cancer Research* 46, no. 9 (1986): 4372-78.

[36] Portier, C.J. "Design of Two-Year Carcinogenicity Experiments: Dose Allocation, Animal Allocation and Sacrifice Times." In *Statistics in Toxicology*, edited by D. Krewski and C. Franklin, 457âĂŞ69. Gordon & Breach, New York, 1991.

[37] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

[38] Rahman, M.A., and K.K. Lin. "A Comparison of False Positive Rates of Peto and poly-3 Methods for Long-Term Carcinogenicity Data Analysis Using Multiple Comparison Adjustment Method Suggested by Lin and Rahman." *Journal of Biopharmaceutical Statistics* 18, no. 5 (2008): 949-58, doi:10.1080/10543400802287628.

[39] Shanno, D.F. "Conditioning of Quasi-Newton Methods for Function Minimization." *Mathematics of Computation* 24, no. 111 (1970): 647-56, doi:10.2307/2004840.

[40] Tarone, R.E. "Tests for Trend in Life Table Analysis." *Biometrika* 62, no. 3 (1975): 679-82, doi:10.2307/2335528.

[41] Wilson, M.P., and M.R. Schwarzman. "Toward a New U.S. Chemicals Policy: Rebuilding the Foundation to Advance New Science, Green Chemistry, and Environmental Health." *Environ Health Perspect* 117 (2009): 1202-1209, doi:10.1289/ehp.0800404.

[42] Woodruff, R.S. "A Simple Method for Approximating the Variance of a Complicated Estimate." *Journal of the American Statistical Association* 66, no. 334 (1971): 411, doi:10.2307/2283947.

Appendix A

SUPPLEMENTAL MATERIALS FOR CHAPTER 1

A.1 Bieler & Williams Variance Calculation

Bieler & Williams [5] use the delta-method to derive the form of $Var(p_i^*)$ for the poly-k test under the null hypothesis; the resulting estimator accounts for the variability introduced by the number-at-risk quantity n_i^*. To obtain a more tractable, equivalent, variance estimator, they utilize a delta-method procedure introduced by Woodruff (1971) [42]. Finally, for stability, they pool variance estimates across groups to obtain the final form of the variance used in their ratio trend test version of the poly-k.

Throughout, we will be using the following notation:
$n = \sum_i n_i$, where n_i = number of subjects in group i
y_{ij} = tumor status (0 or 1) for animal j in group i
y_i = number in group i with tumor present at death
$\alpha_{ij} = \left[(t_{ij}/t_{max})^k\right]^{\mathbb{1}_{y_{ij}=0}}$, for k chosen based on Weibull assumption
$n_i^* = \sum_{j=1}^{n_i} \alpha_{ij}$
$p_i^* = y_i/n_i^*$
$p^* = \sum_i y_i / \sum_i n_i^*$ and $p = \sum_i y_i / \sum_i n_i$

A.1.1 Delta-method

For any treatment group i, we have the asymptotic result:

$$\sqrt{n_i}\begin{bmatrix} \frac{y_i}{n_i} & - & p \\ \frac{n_i^*}{n_i} & - & E\left[\frac{n_i^*}{n_i}\right] \end{bmatrix} \longrightarrow_d N\left(\vec{0} \quad , \quad \begin{bmatrix} Var\left(\frac{y_i}{\sqrt{n_i}}\right) & Cov\left(\frac{y_i}{\sqrt{n_i}}, \frac{n_i^*}{\sqrt{n_i}}\right) \\ Cov\left(\frac{y_i}{\sqrt{n_i}}, \frac{n_i^*}{\sqrt{n_i}}\right) & Var\left(\frac{n_i^*}{n_i}\right) \end{bmatrix} \right)$$

Now select $g(x,y) = \frac{x}{y}$ and apply the δ-method:

$$\begin{aligned} &\sqrt{n_i}\left[g\left(\tfrac{y_i}{n_i}, \tfrac{n_i^*}{n_i}\right) - g\left(p, E\left[\tfrac{n_i^*}{n_i}\right]\right)\right] \\ &\longrightarrow_d N\left(\vec{0} \quad , \quad \left[\tfrac{\partial g}{\partial p_i}, \tfrac{\partial g}{\partial\left(n_i^*/n_i\right)}\right] \begin{bmatrix} Var\left(\frac{y_i}{\sqrt{n_i}}\right) & Cov\left(\frac{y_i}{\sqrt{n_i}}, \frac{n_i^*}{\sqrt{n_i}}\right) \\ Cov\left(\frac{y_i}{\sqrt{n_i}}, \frac{n_i^*}{\sqrt{n_i}}\right) & Var\left(\frac{n_i^*}{n_i}\right) \end{bmatrix} \begin{bmatrix} \partial g/\partial p_i \\ \partial g/\partial\left(\frac{n_i^*}{n_i}\right) \end{bmatrix} \right) \end{aligned}$$

From this result,

$$\begin{aligned} Var(p_i^*) &= E\left[\frac{\partial g}{\partial y_i}(y_i - n_i p) + \frac{\partial g}{\partial n_i^*}(n_i^* - E n_i^*)\right]^2, \text{ where } \frac{\partial g}{\partial p_i} = \frac{n_i}{n_i^*} \text{ and } \frac{\partial g}{\partial(n_i^*/n_i)} = \frac{-y_i n_i}{n_i^{*2}} \\ &= \frac{1}{n_i^{*2}}\left[Var(y_i) + p_i^{*2} Var(n_i^*) - 2p_i^* Cov(y_i, n_i^*)\right] \end{aligned}$$

This expression agrees with equation (4) of Bieler & Williams [5]. If $n_i^* = n_i$, then this reduces to $\frac{Var(y_i)}{n_i^2} = \frac{n_i pq}{n_i^2} = \frac{pq}{n_i}$, i.e. the usual variance estimate when n_i is not a random variable. The authors note that this expression can be made more tractable and appeal to a result by Woodruff [42] to obtain a more data-based formula.

A.1.2 Alternative form of delta-method (Woodruff 1971)

Woodruff [42] applies the δ-method to an example involving a ratio, but inverts the summations involved. We will first illustrate his method using his own notation (simplifying his example to match the Bieler & Williams situation) and then apply the result directly to the situation covered in the previous section.

Woodruff assumes a stratified random sampling situation (without replacement) from a finite source population (taking sample n_i from each stratum i of size N_i), and his calculations weight the variables by the inverse of the sampling probability. (In our case, the population is infinite and we will ignore this weighting.) His design collects data on l draw-variates on $j = 1...n_i$ animals in $i = 1...I + 1$ strata; since $l \in \{1, 2\}$ in our case, we will reduce his formulas to 2 draw-variates. Here, $X_{1ij} = \frac{x_{1ij}}{n_i/N_i}$ is the first random weighted draw-variate; it corresponds to the count data y_{ij} in Bieler & Williams' example. $X_{2ij} = \frac{x_{2ij}}{n_i/N_i}$ is the second weighted draw-variate, corresponding to Bieler & Williams' α_{ij}. Finally, $E(X_{lij}) = Y_{lij}$.

The δ-method can be expressed as follows: If we consider a function F that takes as arguments the random draw-variates X_{lij} (subscripts as defined above), and we further define the quantity

$$F' = \sum_{l=1}^{2} \sum_{j=1}^{n_i} \frac{\partial F}{\partial X_{lij}} X_{lij}, \text{ where } \frac{\partial F}{\partial X_{lij}} \text{ evaluated at expectations}$$

then $VarF' = E[F' - EF']^2$ is exactly the form of the variance of F from the δ-method. Woodruff's contribution is to rewrite $F' = \sum_{j=1}^{n_i} \sum_{l=1}^{2} \frac{\partial F}{\partial X_{lij}} X_{lij}$ and substitute data-based estimates for expectations, to obtain an alternate, equivalent, variance computation. The usual δ-method variance calculation takes a *function of average contributions* from the observations; turning it around yields an *average of the function of contributions* from each observation, and we calculate a variance estimate using an empirical sum of contributions from each subject. His example is worked below.

Step 1 Define $F = \frac{\sum_{j=1}^{n_i} X_{1ij}}{\sum_{j=1}^{n_i} X_{2ij}}$. So $\frac{\partial F}{\partial X_{1ij}} = \frac{1}{\sum_{j=1}^{n_i} X_{2ij}}$ and $\frac{\partial F}{\partial X_{2ij}} = \frac{-\sum_{j=1}^{n_i} X_{1ij}}{\left(\sum_{j=1}^{n_i} X_{2ij}\right)^2}$.

Step 2 Calculate $F' = \sum_{l=1}^{2} \sum_{j=1}^{n_i} \frac{\partial F}{\partial X_{lij}} X_{lij} = \sum_{j=1}^{n_i} \left[\frac{X_{1ij}}{\sum_{j=1}^{n_i} E[X_{2ij}]}\right] + \sum_{j=1}^{n_i} \left[\frac{-X_{2ij} \sum_{j=1}^{n_i} E[X_{1ij}]}{\left(\sum_{j=1}^{n_i} E[X_{2ij}]\right)^2}\right]$.

Step 3 Flip around the summation:
$F' = \sum_{j=1}^{n_i} \left[\frac{X_{1ij}}{\sum_{j=1}^{n_i} E[X_{2ij}]} - \frac{X_{2ij} \sum_{j=1}^{n_i} E[X_{1ij}]}{\left(\sum_{j=1}^{n_i} E[X_{2ij}]\right)^2} \right] = \sum_{j=1}^{n_i} \left[\frac{X_{1ij}}{Y_{2i.}} - \frac{X_{2ij} Y_{1i.}}{Y_{2i.}^2} \right] = \sum_{j=1}^{n_i} U_{ij}.$

Step 4 Since this is formulated as a random sampling situation without replacement, we apply the usual finite population correction factor to obtain $Var(F')$:
$E(F' - EF')^2 = \frac{N_i - n_i}{N_i} \frac{1}{n_i} Var(\sum_{j=1}^{n_i} U_{ij}) = \frac{N_i - n_i}{N_i} \frac{1}{n_i} n_i^2 Var(U_{ij}) = \frac{N_i - n_i}{N_i} n_i \frac{\sum_{j=1}^{N_i} (V_{ij} - \bar{V_i})^2}{N_i - 1}$
(substituting population average for $Var(U_{ij})$)
V_{ij} refers to the population values that correspond to the random variables U_{ij}. Specifically: $V_{ij} = \frac{N_i}{n_i} \left[\frac{Y_{1ij}}{Y_{2i.}} - \frac{Y_{2ij} Y_{1i.}}{Y_{2i.}^2} \right]$ and $\bar{V_i} = \frac{N_i}{n_i} \left[\frac{\bar{Y_{1i}}}{Y_{2i.}} - \frac{\bar{Y_{2i}} Y_{1i.}}{Y_{2i.}^2} \right]$.

Step 5 Finally, we substitute the V-value estimates for the expectation expressions.

$$\begin{aligned} Var(F') &\approx \frac{N_i - n_i}{N_i} \frac{n_i}{N_i - 1} \sum_{j=1}^{N_i} \left[\frac{N_i}{n_i} \left(\frac{Y_{1ij}}{Y_{2i.}} - \frac{Y_{2ij} Y_{1i.}}{Y_{2i.}^2} - \frac{\bar{Y_{1i}}}{Y_{2i.}} + \frac{\bar{Y_{2i}} Y_{1i.}}{Y_{2i.}^2} \right) \right]^2 \\ &= \frac{N_i - n_i}{N_i} \frac{n_i}{N_i - 1} \frac{N_i^2}{n_i^2} \frac{1}{Y_{2i.}^2} \sum_{j=1}^{N_i} \left[\left(Y_{1ij} - \frac{Y_{2ij} Y_{1i.}}{Y_{2i.}} \right) - \left(\bar{Y_{1i}} - \frac{\bar{Y_{2i}} Y_{1i.}}{Y_{2i.}} \right) \right]^2 \end{aligned}$$

Such a calculation follows analogously in the Bieler & Williams case. We note that in the final variance expression, $\frac{N_i - n_i}{N_i} \to 1$ since $N_i \to \infty$. As noted above, we will also ignore the

part of this expression ($\frac{N_i^2}{n_i^2}$) that comes from the sample weighting.

$$\begin{aligned} F &= \frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{j=1}^{n_i} \alpha_{ij}} = \frac{y_i}{n_i^*} = p_i^* \\ F' &= \sum_{j=1}^{n_i} \left(\frac{y_{ij}}{n_i^*} - \frac{y_i \alpha_{ij}}{(n_i^*)^2} \right) = \sum_{j=1}^{n_i} \left(\frac{y_{ij}}{n_i^*} - \frac{p_i^* \alpha_{ij}}{n_i^*} \right) \\ Var(F') &\approx \frac{n_i}{n_i - 1} \sum_{j=1}^{n_i} \left(\frac{y_{ij} - p_i^* \alpha_{ij}}{n_i^*} - \frac{\frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - p_i^* \alpha_{ij})}{n_i^*} \right)^2 \\ &= \frac{n_i}{n_i - 1} \sum_{j=1}^{n_i} (z_{ij} - \bar{z_i})^2 \end{aligned} \tag{A.1}$$

This gives us the estimate for $Var(p_i^*)$ expressed by equations (5) and (6) of Bieler & Williams [5]. This is a consistent variance estimate, but it turns out not to be sufficiently stable. To correct for this, Bieler & Williams pooled the variance estimate across the $I + 1$ treatment groups.

A.1.3 Pooled variance estimator

The final form of Bieler & Williams' variance estimator is a pooled variance based on the estimate (A.1) derived above. If we assume, under the null, that $p_i^* = p^*$ is the same across the treatment groups, and define $r_{ij} = n_i^* z_{ij}$, then we can express:

$$Var(p_i^*) \approx \frac{n_i}{n_i - 1} \sum_j (z_{ij} - \bar{z_i})^2 = \frac{n_i}{n_i^{*2}} \frac{1}{n_i - 1} \sum_j (r_{ij} - \bar{r_i})^2 = \frac{n_i}{n_i^{*2}} \sigma_i^2$$

Pooling σ_i^2 over the $I + 1$ treatment groups, we obtain:

$$Var_{\text{pooled}}(p_i^*) = \frac{n_i}{n_i^{*2}} \frac{\sum_{i=0}^{I} (n_i - 1) \sigma_i^2}{\sum_{i=0}^{I} (n_i - 1)} = \frac{n_i}{n_i^{*2}} \frac{\sum_{i=0}^{I} \sum_{j=1}^{n_i} (r_{ij} - \bar{r_i})^2}{n - (I + 1)}$$

This is the final form used in section 1.2.2 for equation (1.2).

Appendix B

SUPPLEMENTAL MATERIALS FOR CHAPTER 2

B.1 Simulation Settings

Table B.1 summarizes the simulation settings used to explore tests' behavior under different underlying tumor hazards. Simulations were run using either Weibull, log-logistic, or Gompertz tumor hazards, at varying parameter settings. For each of these, tests were examined under Ho ($\eta = 0$, corresponding to the first p_1^* in each row) and under Ha ($\eta > 0$, as indicated, corresponding to the second (or third) p_1^* in each row).

Table B.1: Simulation Settings with Weibull, log-logistic, and Gompertz tumor hazards

			Weibull		log-logistic		Gompertz	
p_0 *	p_1 *	η	k	λ †	k	λ †	b	a †
0.15	0.15, 0.3, 0.45	0, 1.195, 2.679	0.5	2.540 x 10^{-4}	0.5	2.994 x 10^{-4}	0.005	1.191 x 10^{-3}
0.15	0.15, 0.3, 0.45	0, 1.195, 2.679	1	1.563 x 10^{-3}	1	1.697 x 10^{-3}	0.01	8.885 x 10^{-4}
0.15	0.15, 0.3, 0.45	0, 1.195, 2.679	1.5	2.864 x 10^{-3}	1.5	3.025 x 10^{-3}	0.015	6.486 x 10^{-4}
0.15	0.15, 0.3, 0.45	0, 1.195, 2.679	2	3.876 x 10^{-3}	2	4.039 x 10^{-3}	0.02	4.640 x 10^{-4}
0.15	0.15, 0.3, 0.45	0, 1.195, 2.679	2.2	4.210 x 10^{-3}	2.2	4.371 x 10^{-3}	0.022	4.038 x 10^{-4}
0.15	0.15, 0.3, 0.45	0, 1.195, 2.679	2.5	4.649 x 10^{-3}	2.5	4.804 x 10^{-3}	0.025	3.260 x 10^{-4}
0.15	0.15, 0.3, 0.45	0, 1.195, 2.679	3	5.247 x 10^{-3}	3	5.393 x 10^{-3}	0.03	2.252 x 10^{-4}
0.15	0.15, 0.3, 0.45	0, 1.195, 2.679	3.5	5.722 x 10^{-3}	3.5	5.858 x 10^{-3}	0.035	1.534 x 10^{-4}
0.15	0.15, 0.3, 0.45	0, 1.195, 2.679	3.8	5.961 x 10^{-3}	3.8	6.091 x 10^{-3}	0.038	1.210 x 10^{-4}
0.15	0.15, 0.3, 0.45	0, 1.195, 2.679	5	6.686 x 10^{-3}	5	6.797 x 10^{-3}	0.05	4.508 x 10^{-5}
0.15	0.15, 0.3, 0.45	0, 1.195, 2.679	10	8.018 x 10^{-3}	10	8.084 x 10^{-3}	0.1	4.95 x 10^{-7}
0.3	0.3, 0.45	0, 0.676	0.5	1.223 x 10^{-3}	0.5	1.766 x 10^{-3}	0.005	2.615 x 10^{-3}
0.3	0.3, 0.45	0, 0.676	1	3.430 x 10^{-3}	1	4.121 x 10^{-3}	0.01	1.950 x 10^{-3}
0.3	0.3, 0.45	0, 0.676	1.5	4.836 x 10^{-3}	1.5	5.466 x 10^{-3}	0.015	1.423 x 10^{-3}
0.3	0.3, 0.45	0, 0.676	2	5.743 x 10^{-3}	2	6.295 x 10^{-3}	0.02	1.018 x 10^{-3}
0.3	0.3, 0.45	0, 0.676	2.2	6.018 x 10^{-3}	2.2	6.542 x 10^{-3}	0.022	8.861 x 10^{-4}
0.3	0.3, 0.45	0, 0.676	2.5	6.366 x 10^{-3}	2.5	6.851 x 10^{-3}	0.025	7.154 x 10^{-4}
0.3	0.3, 0.45	0, 0.676	3	6.819 x 10^{-3}	3	7.249 x 10^{-3}	0.03	4.943 x 10^{-4}
0.3	0.3, 0.45	0, 0.676	3.5	7.162 x 10^{-3}	3.5	7.548 x 10^{-3}	0.035	3.366 x 10^{-4}
0.3	0.3, 0.45	0, 0.676	3.8	7.331 x 10^{-3}	3.8	7.694 x 10^{-3}	0.038	2.656 x 10^{-4}
0.3	0.3, 0.45	0, 0.676	5	7.824 x 10^{-3}	5	8.117 x 10^{-3}	0.05	9.893 x 10^{-5}
0.3	0.3, 0.45	0, 0.676	10	8.673 x 10^{-3}	10	8.834 x 10^{-3}	0.1	1.085 x 10^{-6}

† The second parameter in the tumor hazard is determined according to control group background tumor probability.

* 2-year cumulative background tumor probabilities before competing risks are p_0 for control group and p_1 for highest-dose group

B.2 Uncalibrated Power

Uncalibrated Power was examined to determine selection criteria for an example NTP data set. The simulations reported in Figure B.1 are a subset of those described in Appendix B.1, plus analagous simulations using $p_0 = 0.05$ with $p_1 \in (0.10, 0.15)$.

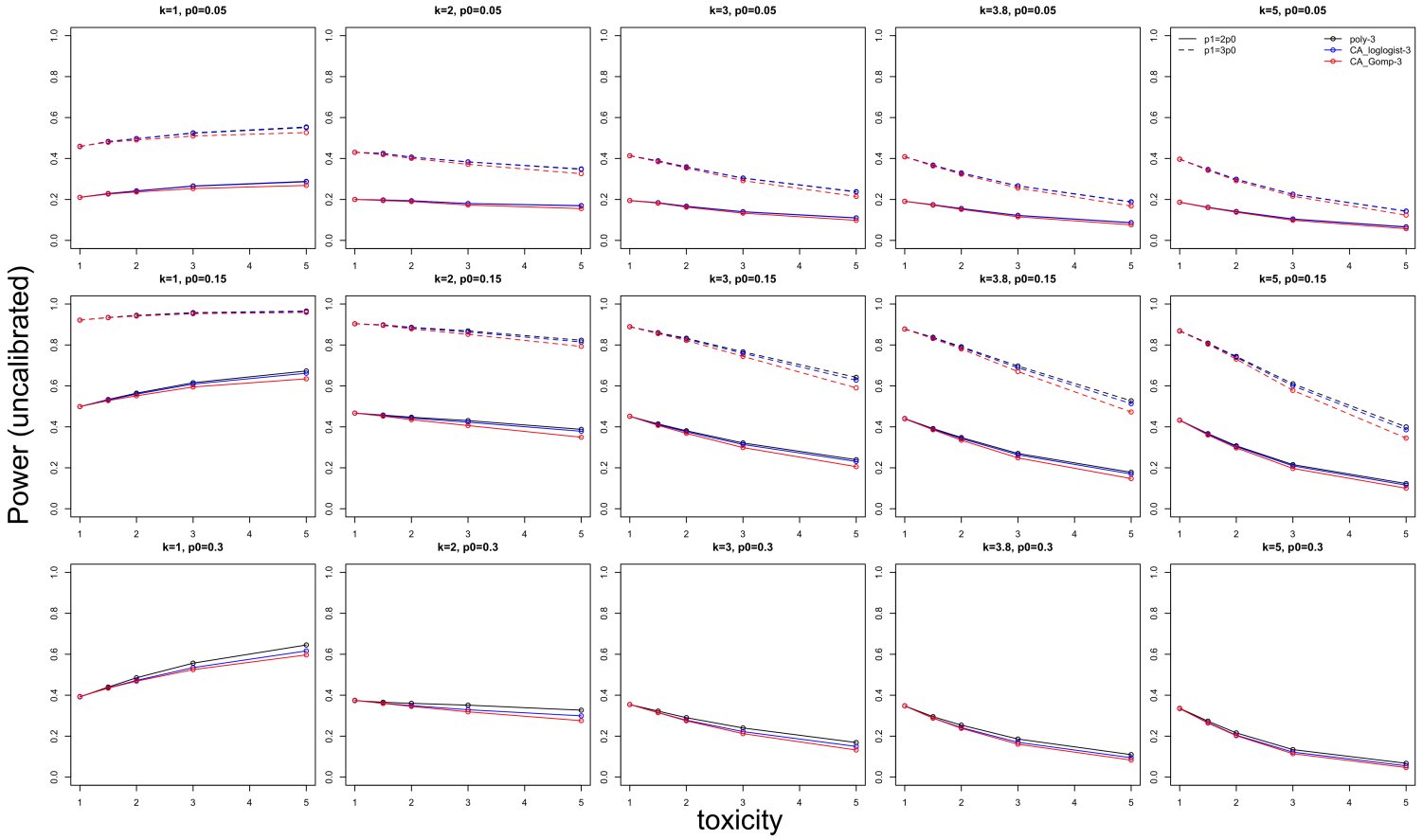


Figure B.1: Data simulated under Ha (un-calibrated Power) with varying p_0 (rows), effect size (line type), and true k_T (columns), for 1-sided tests

B.3 Additional Simulation Results: 1-sided tests

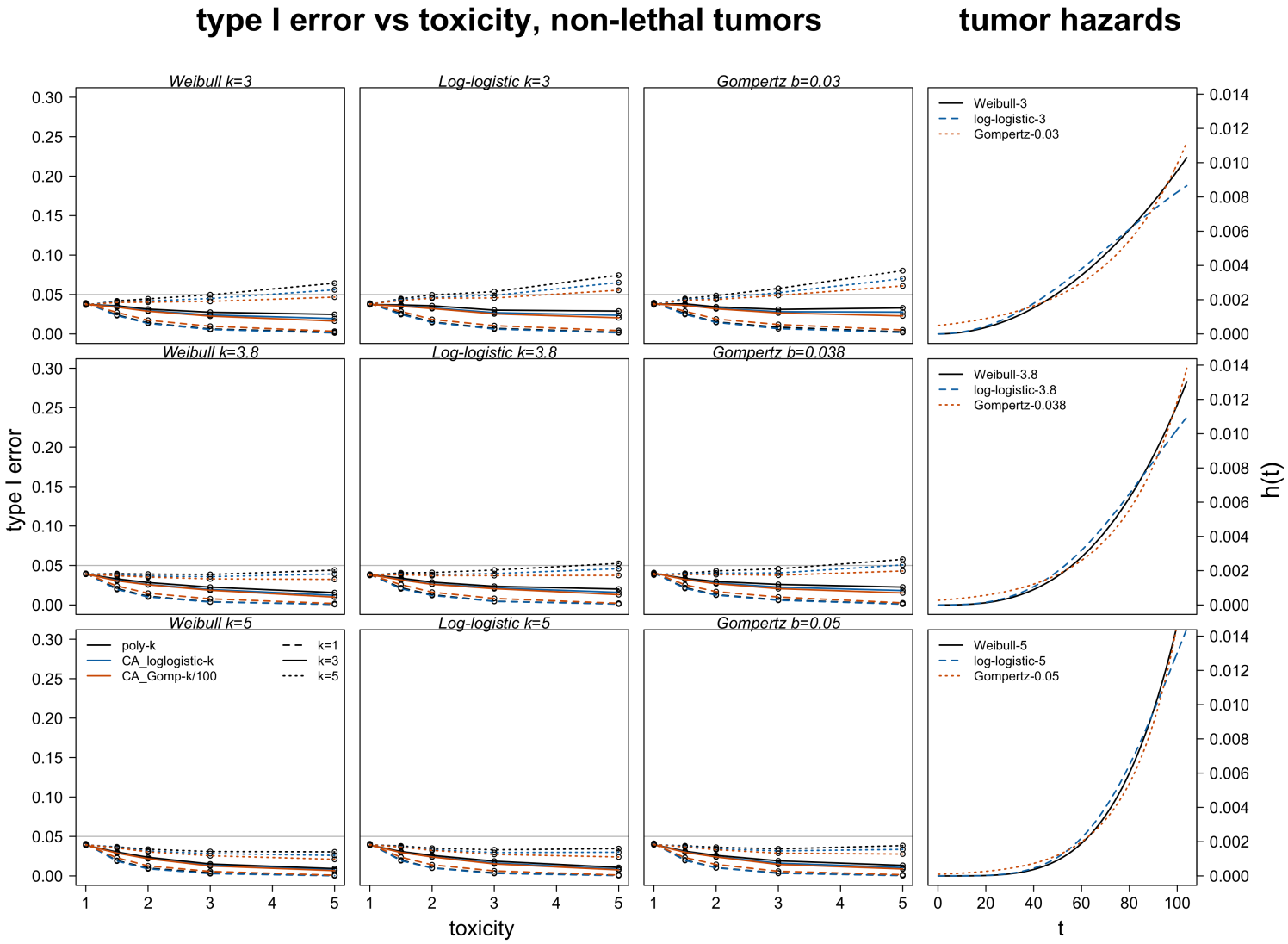


Figure B.2: Data simulated under Ho (type I error) with $p_0 = 0.30$, for nominal 0.05 level 1-sided tests

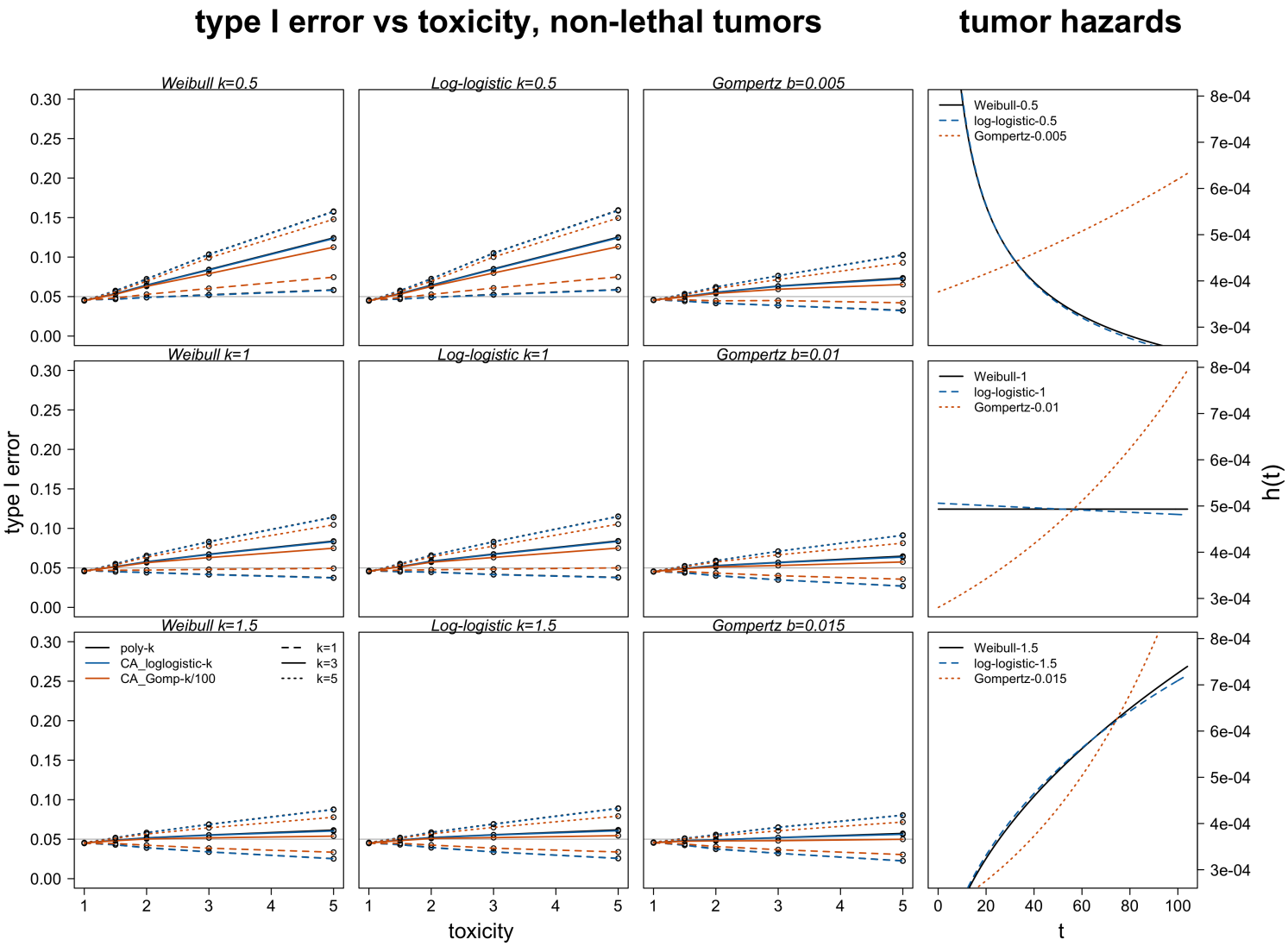


Figure B.3: Data simulated under Ho (type I error) with $p_0 = 0.05$, for nominal 0.05 level 1-sided tests

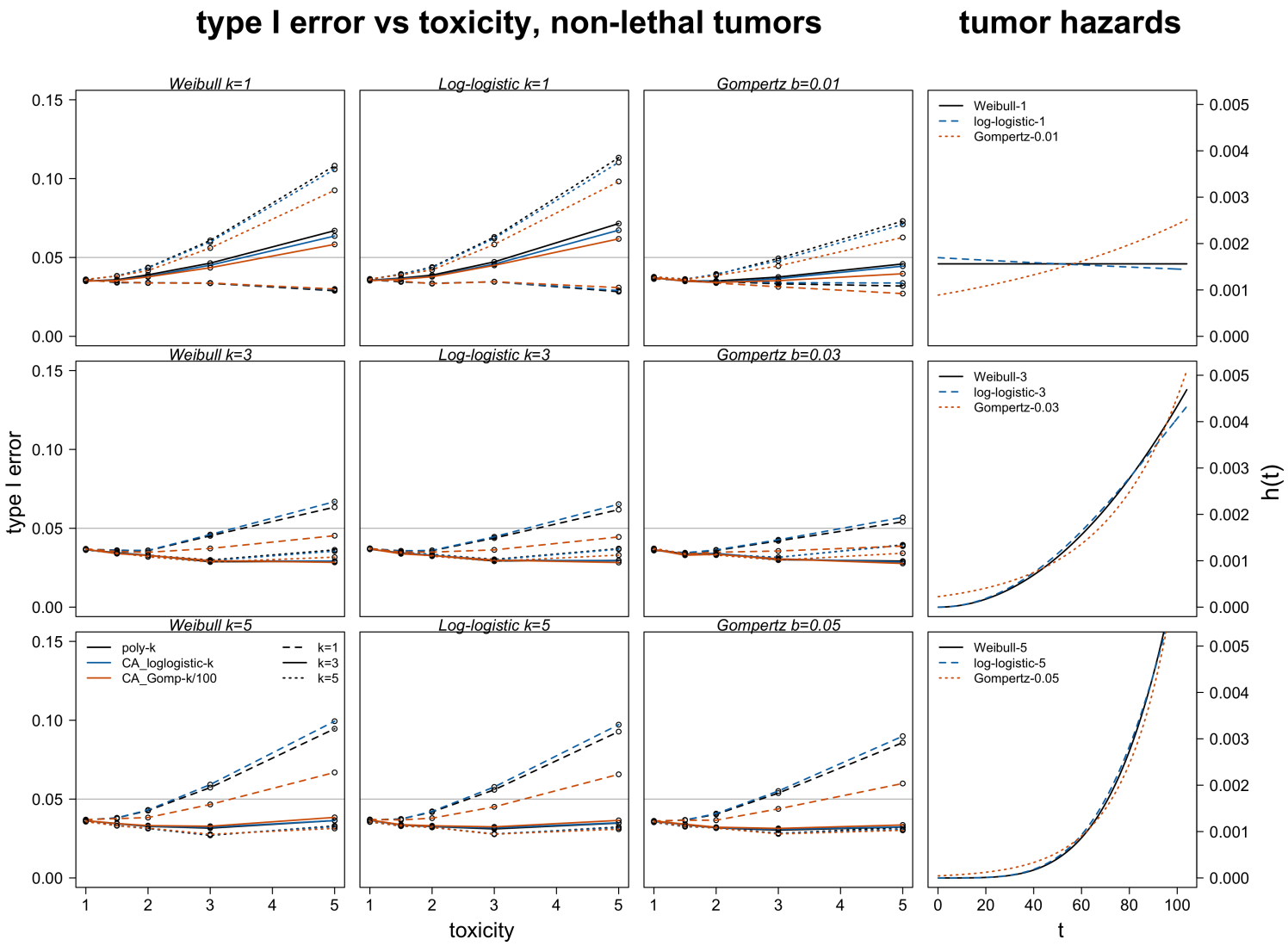


Figure B.4: Data simulated under Ho (type I error) with $p_0 = 0.15$, for nominal 0.05 level 2-sided tests

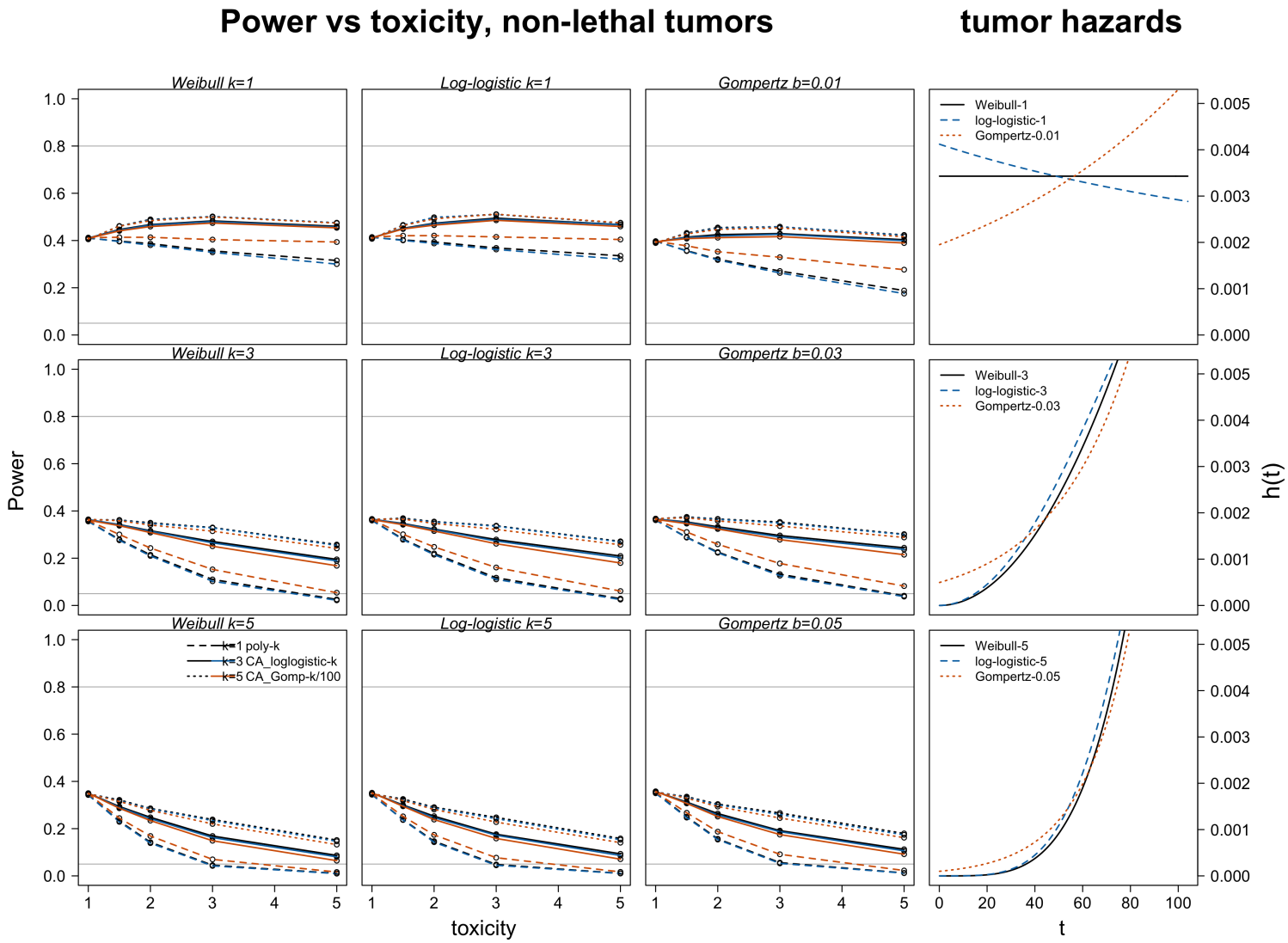


Figure B.5: Data simulated under Ha (Power) with $p_0 = 0.15$, $p_1 = 0.3$, for 2-sided calibrated 0.05 tests

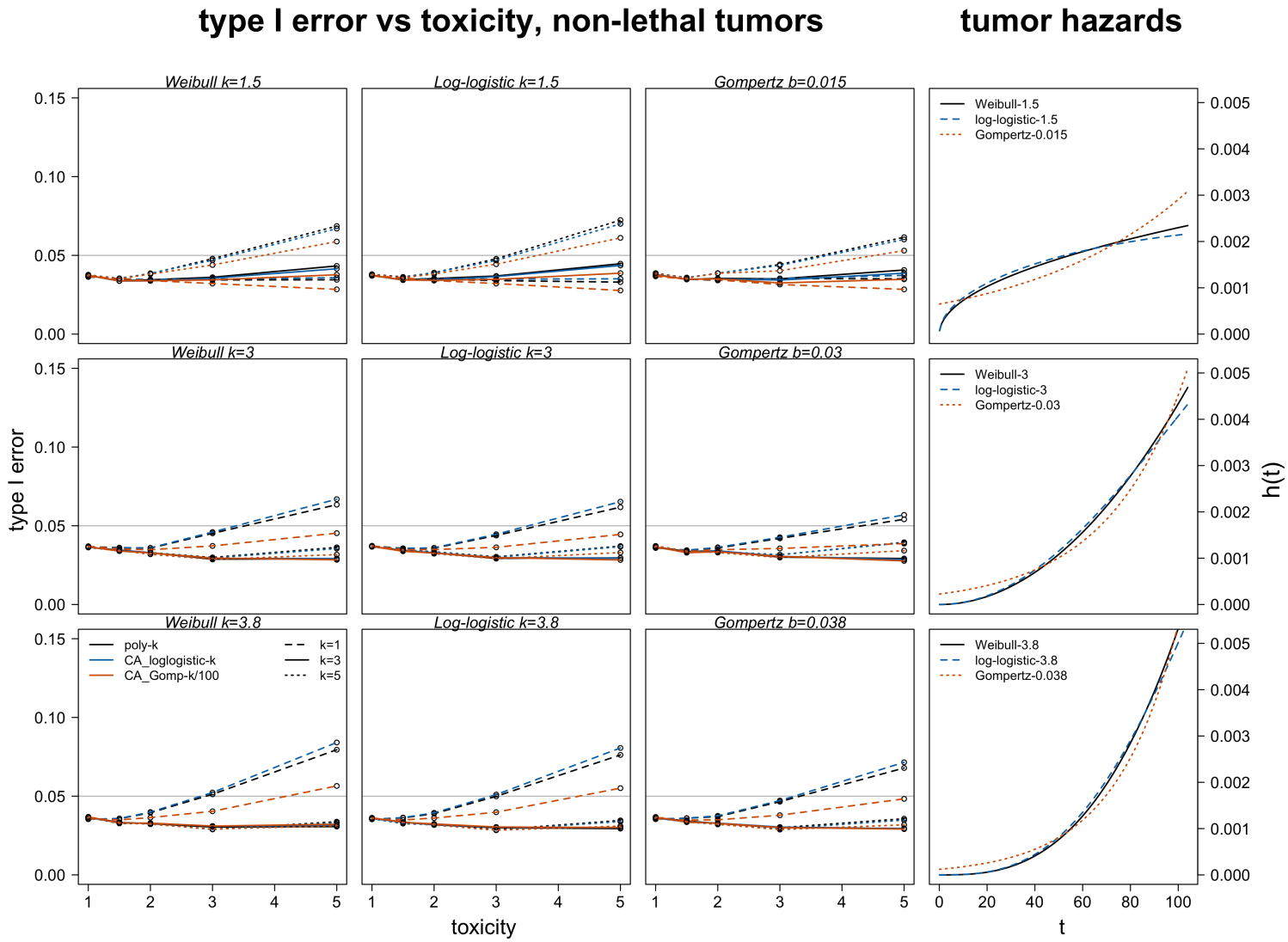


Figure B.6: Data simulated under Ho (type I error) with $p_0 = 0.15$, for nominal 0.05 level 2-sided tests (less extreme departures from poly-3 assumptions)

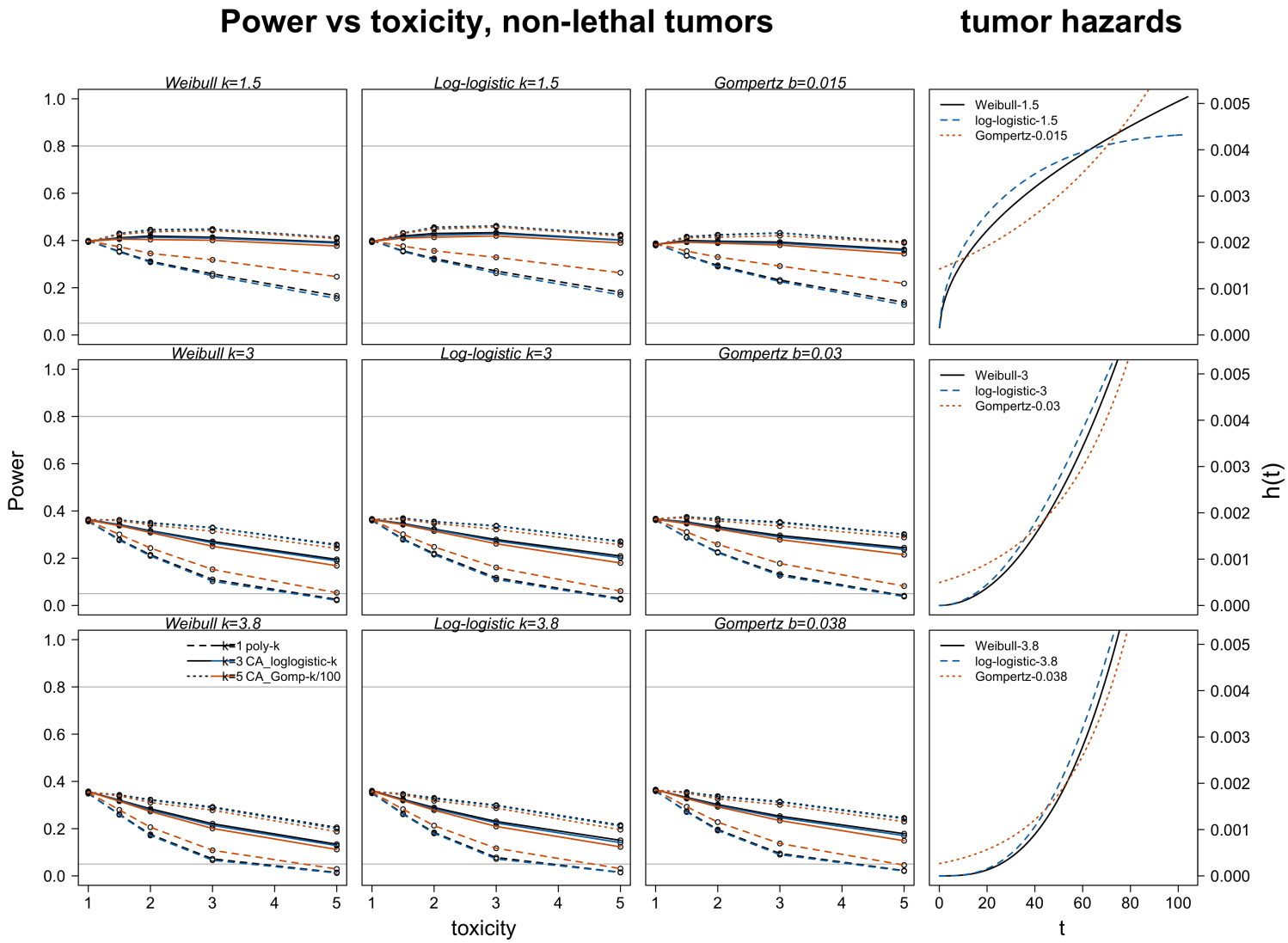


Figure B.7: Data simulated under Ha (Power) with $p_0 = 0.15$, $p_1 = 0.3$, for 2-sided calibrated 0.05 tests (less extreme departures from poly-3 assumptions)

Appendix C

SUPPLEMENTAL MATERIALS FOR CHAPTER 3

C.1 Simulation Settings: Examining the shape of data

C.1.1 How Data were Modeled

Let:

$$
\begin{aligned}
p_1(t) &= P(\text{tumor, avoid nat death by } t|\text{no sacrifice by } t) \\
&= \int_0^t h^T(s) e^{-\int_0^s h^T(u)du - \int_0^t h^{DOC}(u)du - \int_s^t h^{DFT|T}(u,s)du} ds \\
p_2(t) &= P(\text{no tumor, avoid nat death by } t|\text{no sacr by } t) \\
&= e^{-\int_0^t \left(h^{DOC}(u)+h^T(u)du\right)} \\
p_3(t) &= P(\text{tumor by } t, \text{ nat death at } t|\text{no sacrifice by } t) \\
&= \int_0^t \left[h^{DOC}(t) + h^{DFT|T}(t,s)\right] h^T(s) \cdot e^{-\int_0^s h^T(u)du - \int_0^t h^{DOC}(u)du - \int_s^t h^{DFT|T}(u,s)du} ds \\
p_4(t) &= P(\text{no tumor by } t, \text{ nat death at } t|\text{no sacrifice by } t) \\
&= h^{DOC}(t) e^{-\int_0^t \left(h^{DOC}(u)+h^T(u)\right)du} \\
\pi_t &= P(\text{sacrifice at } t|\text{avoid any death before } t) = \frac{n_t}{N \cdot \Pi_{t-1} \cdot (p_1(t)+p_2(t))}; \quad \pi_{tmax} = 1; \\
\Pi_t &= \prod_{s \le t} [1 - \pi_s]
\end{aligned}
$$

Then:

$$\begin{aligned} P(AWT \leq t) &= \Pi_t \cdot p_1(t) \\ P(AWoT \leq t) &= \Pi_t \cdot p_2(t) \\ P(DWT = t) &= \Pi_t \cdot p_3(t) \\ P(DWoT = t) &= \Pi_t \cdot p_4(t) \\ P(SWT = t) &= \Pi_{t-1} \cdot p_1(t) \cdot \pi_t \\ P(SWoT = t) &= \Pi_{t-1} \cdot p_2(t) \cdot \pi_t \end{aligned}$$

(...and $P(DWT \leq t)$ or $P(DWoT \leq t)$ can be obtained by integration)

In simulated data (s_{ij} time to tumor, t_{ij} death time, d_{ij} sacrifice), these correspond to:

$$\begin{aligned} \hat{\mathrm{P}}(AWT \leq t) &= \textstyle\sum_j \mathbb{1}_{s_{ij} \leq t, t_{ij} > t} / N_i \\ \hat{\mathrm{P}}(AWoT \leq t) &= \textstyle\sum_j \mathbb{1}_{s_{ij} > t, t_{ij} > t} / N_i \\ \hat{\mathrm{P}}(DWT \leq t) &= \textstyle\sum_j \mathbb{1}_{s_{ij} \leq t, t_{ij} \leq t, d_{ij} = 0} / N_i \\ \hat{\mathrm{P}}(DWoT \leq t) &= \textstyle\sum_j \mathbb{1}_{s_{ij} > t, t_{ij} \leq t, d_{ij} = 0} / N_i \\ \hat{\mathrm{P}}(SWT = t) &= \textstyle\sum_j \mathbb{1}_{s_{ij} \leq t, t_{ij} \leq t, d_{ij} = 1} / N_i \\ \hat{\mathrm{P}}(SWoT = t) &= \textstyle\sum_j \mathbb{1}_{s_{ij} > t, t_{ij} \leq t, d_{ij} = 1} / N_i \end{aligned}$$

All plots were generated with: group $N = 60$; interim sacrifice number n_t chosen s.t. total interim-sacrificed $\in$ (10 or 12); $L = 1$, $TOX = 1.8$; study-end tumor rate $p_0 = 0.15$ and study-end tumor onset rate in high dose group $= 2p_0$; $k_T \in (1, 1.5, 3, 6)$.

C.1.2 Proportion alive/dead/sacrificed with/without tumor, $k_T = 1$

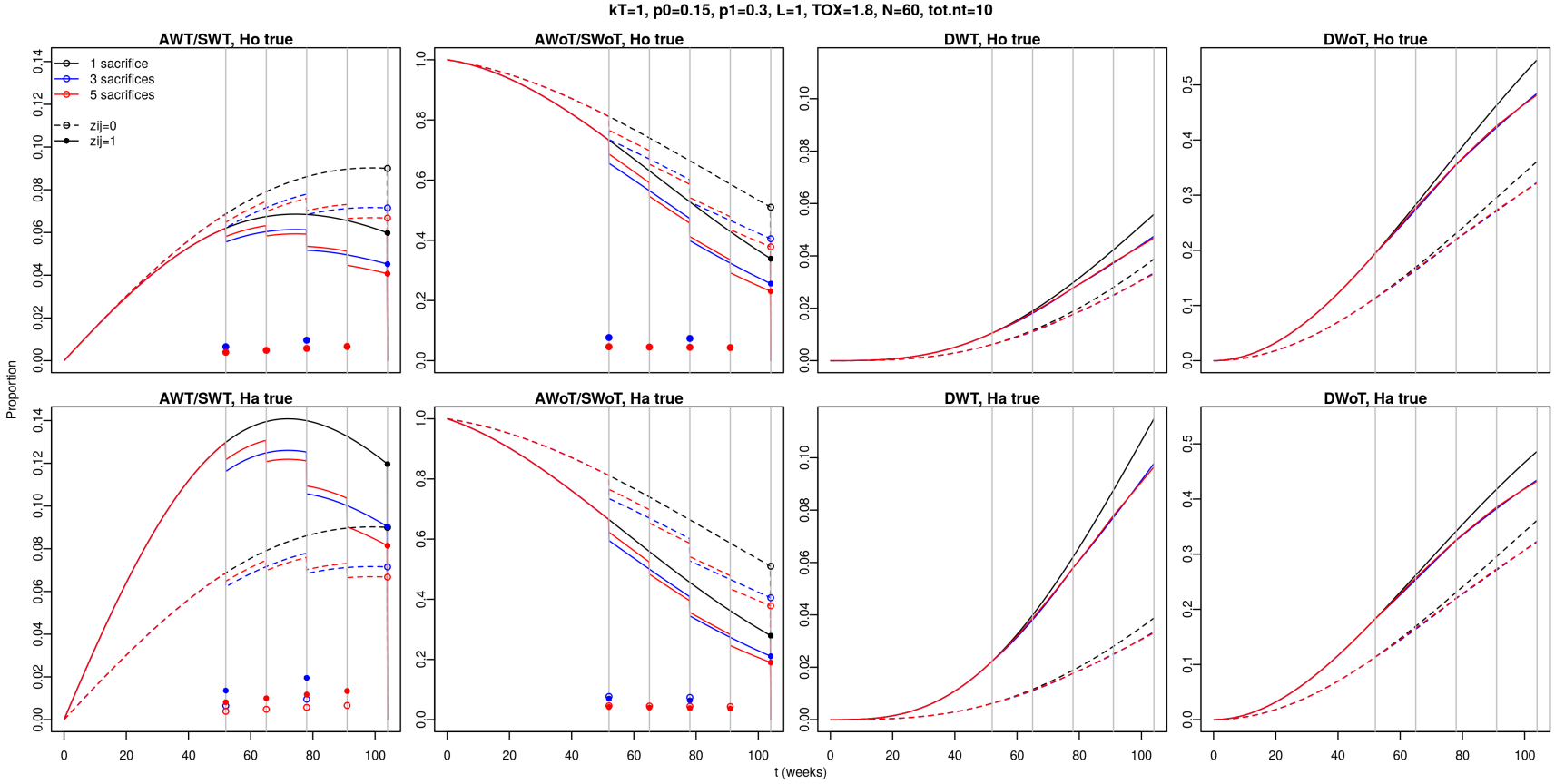


Figure C.1: Simulations under Ho and Ha, with k=1, p_0=0.15, L=1, TOX=1.8, totsacr=10

C.1.3 Proportion alive/dead/sacrificed with/without tumor, $k_T = 1.5$

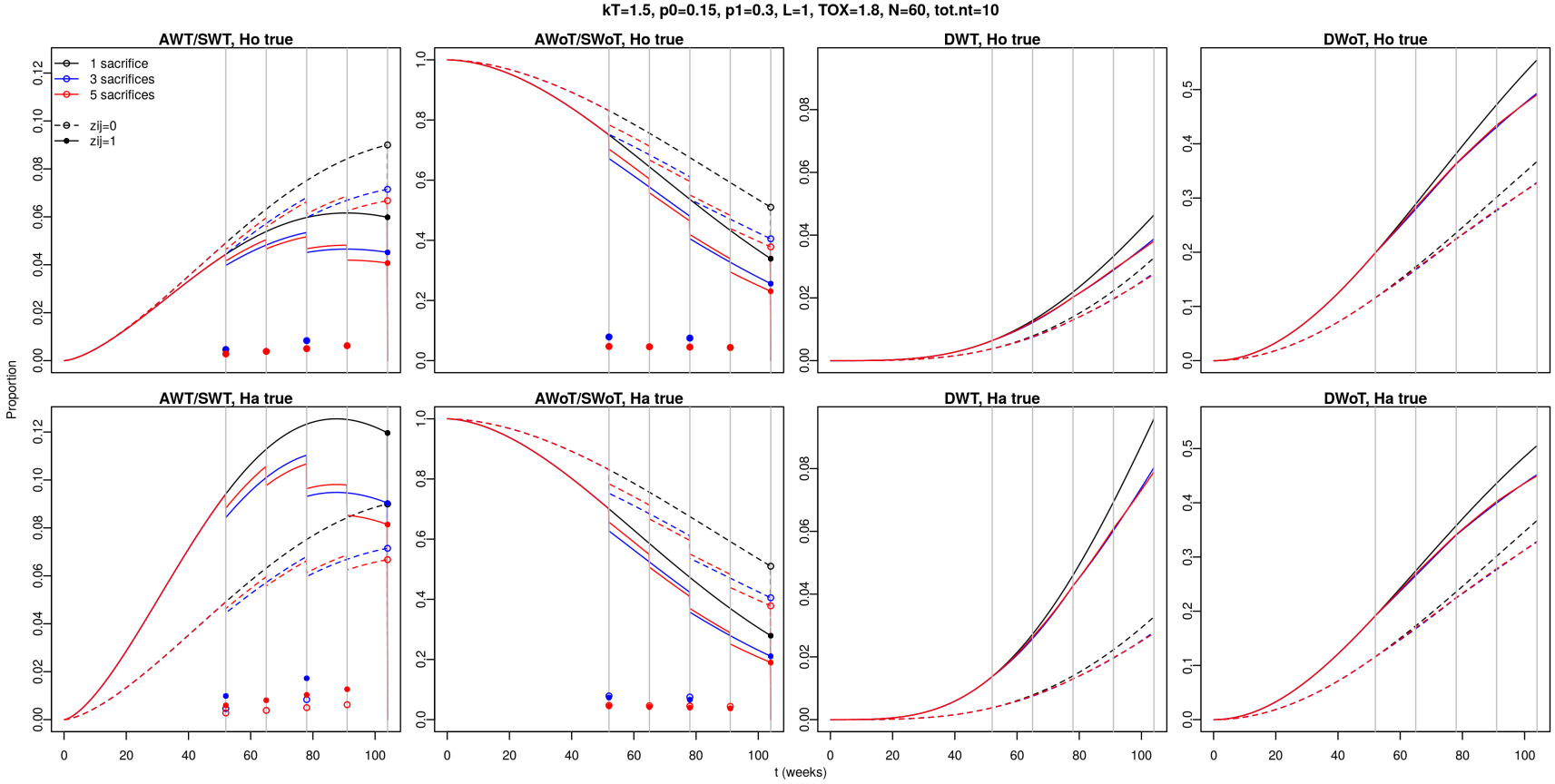


Figure C.2: Simulations under Ho and Ha, with k=1.5, p_0=0.15, L=1, TOX=1.8, totsacr=10

C.1.4 Proportion alive/dead/sacrificed with/without tumor, $k_T = 3$

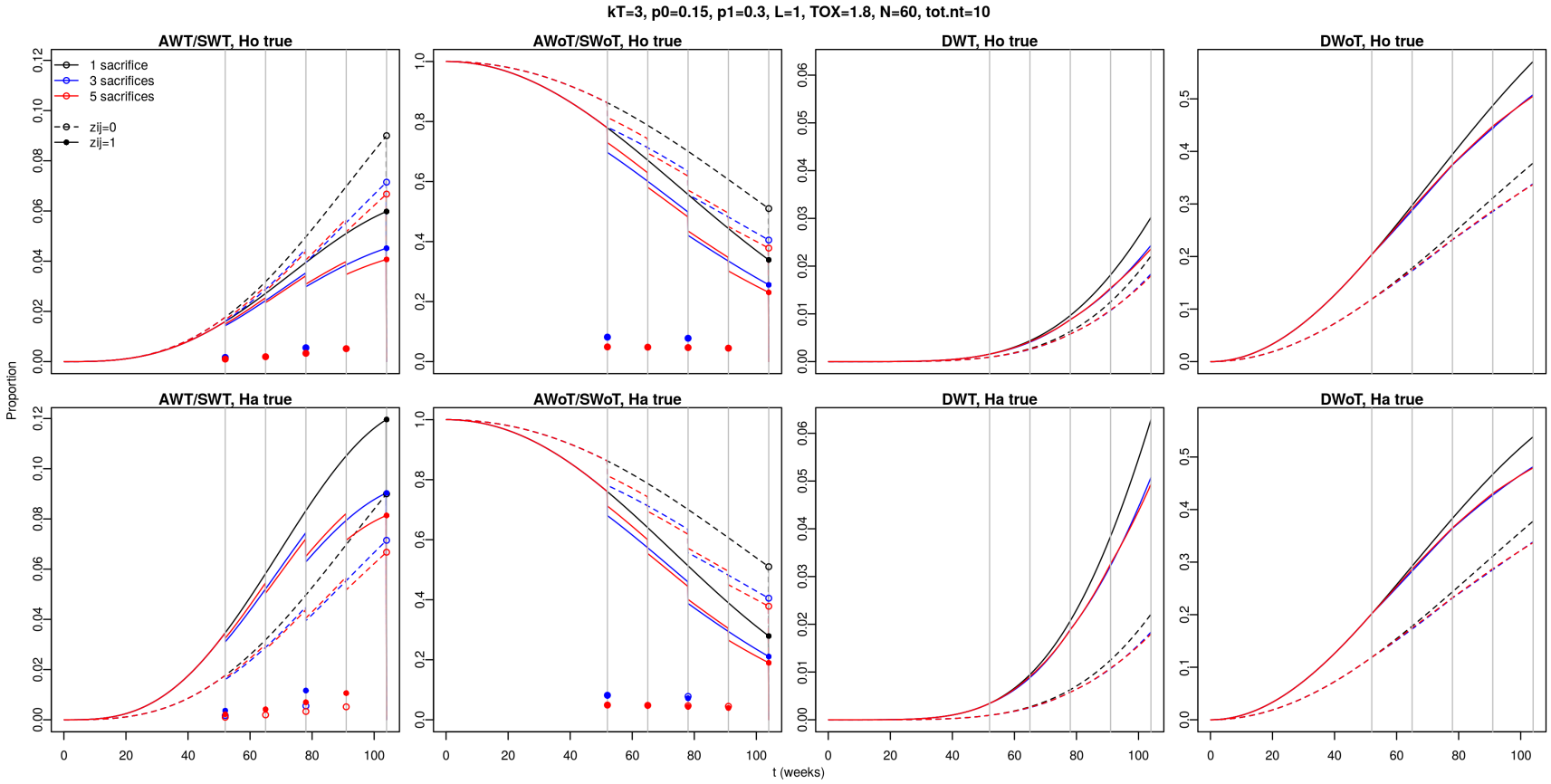


Figure C.3: Simulations under Ho and Ha, with k=3, p_0=0.15, L=1, TOX=1.8, totsacr=10

C.1.5 Proportion alive/dead/sacrificed with/without tumor, $k_T = 6$

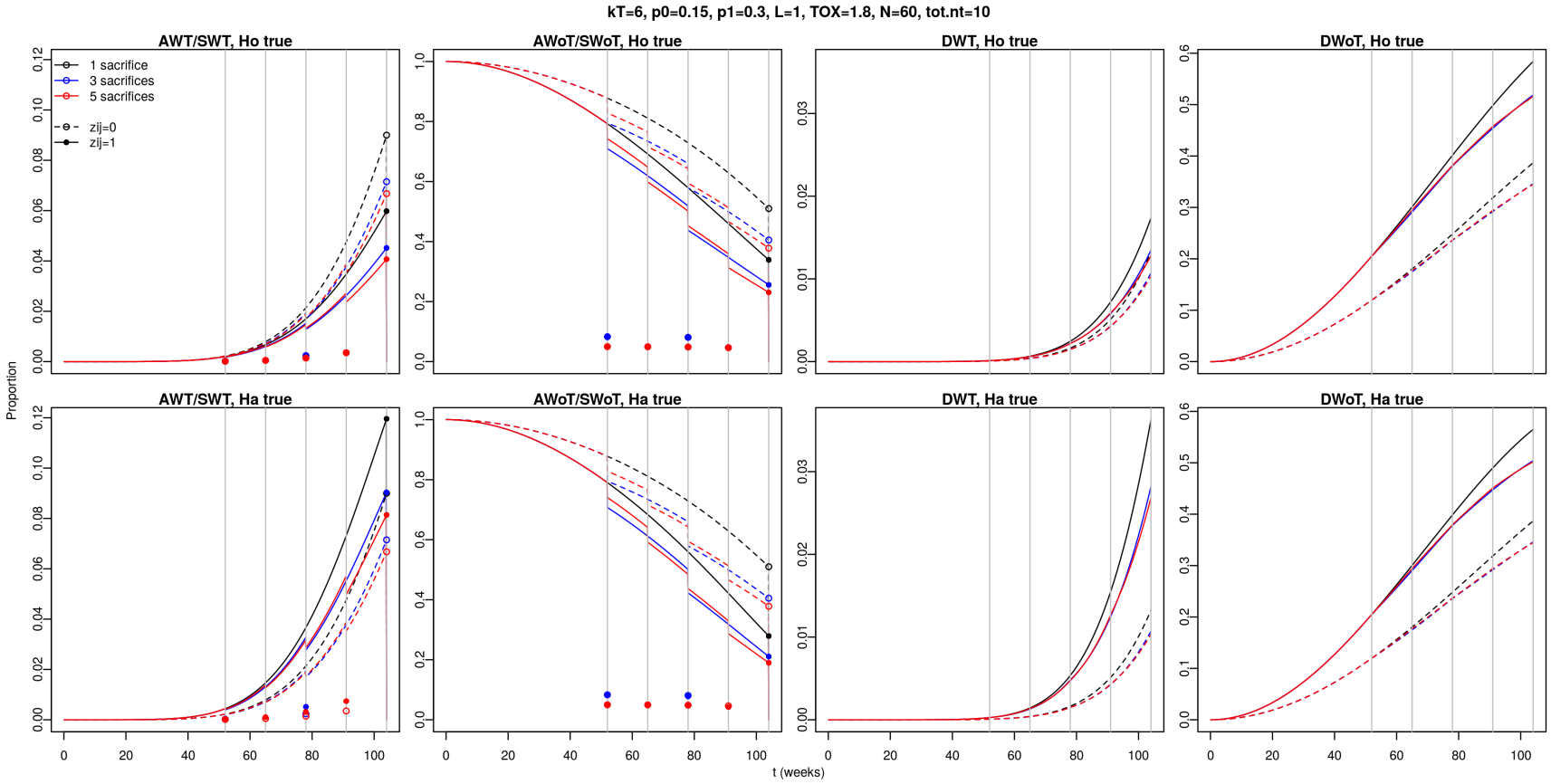


Figure C.4: Simulations under Ho and Ha, with k=6, p_0=0.15, L=1, TOX=1.8, totsacr=10

C.2 Simulation Output: bias & variability

This section contains the data that are represented in figures in the Chapter 3 Results section 3.3.2.

Tables C.1, C.2, C.3, and C.4 present, for varying lethality and toxicity settings, the bias, MSE, and SD for the Moon estimator and MLE-based estimators of $\hat{k_T}$ from those simulations when the cumulative lifetime tumor incidence rate is $\mathbf{p_0 = 0.15}$.

Tables C.5 and C.6 show some of the corresponding comparisons for $\mathbf{p_0 = 0.05}$.

Table C.1: Bias of $\hat{k}_T$ ($p_0 = 0.15$, $L = 1$, $TOX = 1$)

	true k_T	N per group	n sacr	mean and SE for estimates of k					
				Moon			MLE		
				mean	SD	MSE	mean	SD	MSE
Ho True	1	50	1	0	0	1	1.8	1.344	2.447
	1	60	1	0	0	1	1.678	1.114	1.7
	1	60	2	0.608	0.41	0.322	1.334	0.741	0.661
	1	60	3	1.082	0.731	0.54	1.405	0.803	0.809
	1	60	4	1.681	1.171	1.834	1.439	0.835	0.89
	1	60	5	1.808	1.073	1.802	1.419	0.807	0.828
	1.5	50	1	0	0	2.25	2.465	1.697	3.809
	1.5	60	1	0	0	2.25	2.314	1.434	2.717
	1.5	60	2	0.819	0.435	0.653	1.91	1.017	1.202
	1.5	60	3	1.405	0.861	0.75	1.977	1.066	1.365
	1.5	60	4	2.143	1.411	2.406	2.012	1.083	1.436
	1.5	60	5	2.289	1.271	2.239	1.977	1.044	1.317
	3	50	1	0	0	9	4.338	2.446	7.773
	3	60	1	0	0	9	4.122	2.135	5.818
	3	60	2	1.279	0.379	3.106	3.891	2.037	4.942
	3	60	3	2.288	1.102	1.721	3.789	1.842	4.015
	3	60	4	3.532	2.086	4.633	3.777	1.773	3.746
	3	60	5	3.72	1.864	3.994	3.729	1.733	3.534
	6	50	1	0	0	36	6.959	3.032	10.111
	6	60	1	0	0	36	6.698	2.762	8.117
	6	60	2	1.568	0.203	19.682	7.291	2.925	10.222
	6	60	3	3.426	1.113	7.864	7.063	2.918	9.642
	6	60	4	5.951	3.16	9.987	6.901	2.84	8.878
	6	60	5	6.023	2.736	7.485	7.034	2.938	9.697
Ha True	1	50	1	0	0	1	1.579	1.033	1.401
	1	60	1	0	0	1	1.477	0.839	0.932
	1	60	2	0.597	0.361	0.293	1.23	0.556	0.362
	1	60	3	0.991	0.629	0.396	1.287	0.619	0.466
	1	60	4	1.473	0.947	1.12	1.312	0.647	0.515
	1	60	5	1.604	0.872	1.125	1.299	0.633	0.49
	1.5	50	1	0	0	2.25	2.186	1.254	2.043
	1.5	60	1	0	0	2.25	2.077	1.062	1.46
	1.5	60	2	0.811	0.378	0.617	1.765	0.723	0.592
	1.5	60	3	1.305	0.731	0.572	1.823	0.78	0.713
	1.5	60	4	1.9	1.113	1.398	1.851	0.804	0.77
	1.5	60	5	2.07	1.03	1.385	1.836	0.794	0.743
	3	50	1	0	0	9	3.998	1.938	4.752
	3	60	1	0	0	9	3.825	1.667	3.459
	3	60	2	1.284	0.329	3.054	3.563	1.447	2.411
	3	60	3	2.194	0.957	1.565	3.527	1.347	2.092
	3	60	4	3.21	1.666	2.82	3.538	1.331	2.06
	3	60	5	3.432	1.501	2.44	3.498	1.302	1.943
	6	50	1	0	0	36	6.902	2.712	8.165
	6	60	1	0	0	36	6.634	2.408	6.199
	6	60	2	1.576	0.188	19.603	7.064	2.531	7.536
	6	60	3	3.38	0.993	7.848	6.816	2.46	6.714
	6	60	4	5.491	2.588	6.955	6.684	2.322	5.86
	6	60	5	5.643	2.303	5.429	6.737	2.376	6.186

Table C.2: Bias of $\hat{k}_T$ ($p_0 = 0.15$, $L = 1$, $TOX = 1.8$)

	true k_T	N per group	n sacr	mean and SE for estimates of k					
				Moon			MLE		
				mean	SD	MSE	mean	SD	MSE
Ho True	1	50	1	0	0	1	1.683	1.162	1.818
	1	60	1	0	0	1	1.574	0.972	1.274
	1	60	2	0.568	0.387	0.337	1.321	0.703	0.597
	1	60	3	1.037	0.691	0.479	1.382	0.761	0.725
	1	60	4	1.615	1.083	1.552	1.408	0.778	0.773
	1	60	5	1.758	1.003	1.582	1.395	0.761	0.735
	1.5	50	1	0	0	2.25	2.339	1.475	2.877
	1.5	60	1	0	0	2.25	2.198	1.256	2.064
	1.5	60	2	0.766	0.412	0.709	1.886	0.928	1.01
	1.5	60	3	1.347	0.807	0.675	1.944	0.974	1.145
	1.5	60	4	2.066	1.301	2.012	1.976	0.998	1.224
	1.5	60	5	2.234	1.194	1.964	1.949	0.966	1.135
	3	50	1	0	0	9	4.21	2.242	6.493
	3	60	1	0	0	9	3.996	1.97	4.874
	3	60	2	1.203	0.367	3.362	3.833	1.922	4.388
	3	60	3	2.209	1.052	1.734	3.742	1.753	3.625
	3	60	4	3.435	1.949	3.986	3.738	1.673	3.343
	3	60	5	3.637	1.734	3.412	3.694	1.628	3.13
	6	50	1	0	0	36	7.034	3.032	10.261
	6	60	1	0	0	36	6.679	2.71	7.805
	6	60	2	1.486	0.194	20.412	7.234	2.907	9.973
	6	60	3	3.335	1.081	8.271	7.015	2.847	9.134
	6	60	4	5.825	2.98	8.911	6.903	2.8	8.656
	6	60	5	5.901	2.578	6.655	6.965	2.821	8.888
Ha True	1	50	1	0	0	1	1.323	0.801	0.746
	1	60	1	0	0	1	1.247	0.689	0.536
	1	60	2	0.544	0.346	0.328	1.125	0.511	0.277
	1	60	3	0.911	0.588	0.354	1.154	0.561	0.338
	1	60	4	1.331	0.855	0.84	1.163	0.571	0.352
	1	60	5	1.479	0.802	0.873	1.156	0.563	0.341
	1.5	50	1	0	0	2.25	1.885	1.001	1.151
	1.5	60	1	0	0	2.25	1.799	0.873	0.852
	1.5	60	2	0.744	0.367	0.707	1.647	0.685	0.491
	1.5	60	3	1.206	0.69	0.562	1.671	0.717	0.543
	1.5	60	4	1.738	1.023	1.102	1.683	0.728	0.564
	1.5	60	5	1.914	0.953	1.078	1.673	0.726	0.556
	3	50	1	0	0	9	3.67	1.701	3.343
	3	60	1	0	0	9	3.464	1.414	2.215
	3	60	2	1.2	0.323	3.345	3.368	1.284	1.784
	3	60	3	2.064	0.92	1.722	3.316	1.215	1.576
	3	60	4	2.972	1.518	2.303	3.305	1.199	1.531
	3	60	5	3.226	1.387	1.975	3.29	1.173	1.46
	6	50	1	0	0	36	6.571	2.646	7.326
	6	60	1	0	0	36	6.262	2.31	5.403
	6	60	2	1.494	0.179	20.333	6.734	2.439	6.485
	6	60	3	3.26	0.976	8.461	6.484	2.3	5.521
	6	60	4	5.224	2.394	6.333	6.368	2.185	4.907
	6	60	5	5.401	2.147	4.966	6.431	2.233	5.169

Table C.3: Bias of $\hat{k}_T$ ($p_0 = 0.15$, $L = 1.5$, $TOX = 1$)

	true k_T	N per group	n sacr	mean and SE for estimates of k					
				Moon			MLE		
				mean	SD	MSE	mean	SD	MSE
Ho True	1	50	1	0	0	1	1.592	1.143	1.656
	1	60	1	0	0	1	1.501	0.982	1.215
	1	60	2	0.605	0.408	0.323	1.256	0.698	0.553
	1	60	3	1.044	0.718	0.517	1.308	0.747	0.653
	1	60	4	1.578	1.121	1.59	1.335	0.778	0.717
	1	60	5	1.694	1.028	1.539	1.319	0.758	0.676
	1.5	50	1	0	0	2.25	2.248	1.514	2.851
	1.5	60	1	0	0	2.25	2.123	1.303	2.085
	1.5	60	2	0.818	0.434	0.653	1.826	0.949	1.007
	1.5	60	3	1.374	0.853	0.744	1.88	1.008	1.16
	1.5	60	4	2.048	1.356	2.14	1.909	1.032	1.233
	1.5	60	5	2.186	1.225	1.971	1.879	0.999	1.141
	3	50	1	0	0	9	4.138	2.324	6.696
	3	60	1	0	0	9	3.947	2.027	5.003
	3	60	2	1.279	0.378	3.106	3.777	1.923	4.302
	3	60	3	2.275	1.098	1.732	3.694	1.785	3.666
	3	60	4	3.477	2.067	4.5	3.681	1.72	3.423
	3	60	5	3.655	1.844	3.828	3.635	1.674	3.207
	6	50	1	0	0	36	6.848	3.021	9.844
	6	60	1	0	0	36	6.578	2.729	7.783
	6	60	2	1.568	0.203	19.682	7.177	2.865	9.589
	6	60	3	3.423	1.112	7.875	6.98	2.882	9.265
	6	60	4	5.925	3.147	9.906	6.825	2.805	8.546
	6	60	5	5.996	2.724	7.42	6.955	2.904	9.343
Ha True	1	50	1	0	0	1	1.415	0.912	1.004
	1	60	1	0	0	1	1.348	0.793	0.749
	1	60	2	0.596	0.362	0.294	1.169	0.552	0.333
	1	60	3	0.959	0.624	0.39	1.211	0.607	0.412
	1	60	4	1.378	0.911	0.973	1.229	0.627	0.446
	1	60	5	1.491	0.837	0.942	1.219	0.618	0.43
	1.5	50	1	0	0	2.25	2.013	1.151	1.588
	1.5	60	1	0	0	2.25	1.927	0.994	1.171
	1.5	60	2	0.81	0.378	0.619	1.696	0.705	0.536
	1.5	60	3	1.28	0.73	0.58	1.741	0.76	0.635
	1.5	60	4	1.819	1.087	1.283	1.764	0.781	0.68
	1.5	60	5	1.971	1.009	1.239	1.751	0.775	0.664
	3	50	1	0	0	9	3.817	1.822	3.986
	3	60	1	0	0	9	3.67	1.604	3.022
	3	60	2	1.284	0.328	3.053	3.482	1.386	2.152
	3	60	3	2.185	0.952	1.57	3.447	1.313	1.923
	3	60	4	3.159	1.652	2.753	3.455	1.295	1.885
	3	60	5	3.369	1.484	2.337	3.421	1.277	1.806
	6	50	1	0	0	36	6.776	2.697	7.877
	6	60	1	0	0	36	6.504	2.375	5.893
	6	60	2	1.576	0.188	19.604	6.97	2.501	7.193
	6	60	3	3.378	0.991	7.858	6.748	2.437	6.498
	6	60	4	5.468	2.578	6.929	6.626	2.322	5.782
	6	60	5	5.614	2.295	5.417	6.658	2.355	5.976

Table C.4: Bias of $\hat{k}_T$ ($p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$)

	true k_T	N per group	n sacr	mean and SE for estimates of *k*					
				Moon			MLE		
				mean	SD	MSE	mean	SD	MSE
Ho True	1	50	1	0	0	1	1.522	1.008	1.288
	1	60	1	0	0	1	1.429	0.877	0.954
	1	60	2	0.564	0.386	0.339	1.244	0.676	0.516
	1	60	3	0.995	0.677	0.458	1.289	0.721	0.604
	1	60	4	1.511	1.028	1.319	1.31	0.737	0.639
	1	60	5	1.642	0.959	1.332	1.3	0.724	0.615
	1.5	50	1	0	0	2.25	2.144	1.325	2.17
	1.5	60	1	0	0	2.25	2.027	1.137	1.569
	1.5	60	2	0.764	0.412	0.711	1.807	0.891	0.888
	1.5	60	3	1.317	0.804	0.679	1.85	0.933	0.992
	1.5	60	4	1.97	1.253	1.791	1.874	0.957	1.055
	1.5	60	5	2.129	1.154	1.727	1.854	0.931	0.992
	3	50	1	0	0	9	4.035	2.133	5.619
	3	60	1	0	0	9	3.827	1.878	4.211
	3	60	2	1.203	0.365	3.361	3.719	1.817	3.816
	3	60	3	2.193	1.049	1.751	3.642	1.687	3.257
	3	60	4	3.367	1.921	3.825	3.637	1.621	3.034
	3	60	5	3.563	1.711	3.245	3.601	1.589	2.884
	6	50	1	0	0	36	6.897	2.991	9.75
	6	60	1	0	0	36	6.55	2.659	7.372
	6	60	2	1.486	0.193	20.413	7.13	2.863	9.471
	6	60	3	3.331	1.08	8.291	6.918	2.789	8.62
	6	60	4	5.799	2.973	8.878	6.829	2.771	8.365
	6	60	5	5.874	2.569	6.614	6.887	2.795	8.599
Ha True	1	50	1	0	0	1	1.192	0.754	0.606
	1	60	1	0	0	1	1.135	0.666	0.461
	1	60	2	0.543	0.346	0.329	1.058	0.517	0.271
	1	60	3	0.886	0.58	0.35	1.071	0.558	0.316
	1	60	4	1.26	0.826	0.749	1.077	0.569	0.329
	1	60	5	1.386	0.773	0.746	1.074	0.563	0.322
	1.5	50	1	0	0	2.25	1.736	0.929	0.918
	1.5	60	1	0	0	2.25	1.674	0.852	0.756
	1.5	60	2	0.743	0.366	0.707	1.573	0.67	0.454
	1.5	60	3	1.183	0.683	0.566	1.585	0.71	0.511
	1.5	60	4	1.672	0.998	1.024	1.591	0.719	0.524
	1.5	60	5	1.827	0.934	0.98	1.584	0.718	0.522
	3	50	1	0	0	9	3.505	1.641	2.946
	3	60	1	0	0	9	3.322	1.374	1.991
	3	60	2	1.2	0.323	3.344	3.288	1.236	1.61
	3	60	3	2.055	0.914	1.728	3.235	1.189	1.468
	3	60	4	2.927	1.493	2.234	3.225	1.176	1.432
	3	60	5	3.172	1.372	1.911	3.212	1.151	1.37
	6	50	1	0	0	36	6.438	2.594	6.92
	6	60	1	0	0	36	6.124	2.261	5.125
	6	60	2	1.494	0.179	20.335	6.628	2.407	6.189
	6	60	3	3.258	0.975	8.471	6.398	2.282	5.364
	6	60	4	5.207	2.389	6.334	6.293	2.171	4.798
	6	60	5	5.376	2.14	4.969	6.356	2.218	5.046

Table C.5: Bias of $\hat{k}_T$ ($p_0 = 0.05$, $L = 1$, $TOX = 1.8$)

	true k_T	N per group	n sacr	mean and SE for estimates of k						
				Moon			MLE			
				mean	SD	MSE	mean	SD	MSE	prop converged
Ho True	1	60	1	0	0	1	2.625	2.41	8.447	1
	1	60	2	0.715	0.572	0.409	2.063	2.1	5.542	1
	1	60	3	1.587	1.193	1.768	2.131	2.068	5.554	0.9999
	1	60	4	2.807	2.379	8.924	2.187	2.105	5.841	1
	1	60	5	2.883	2.075	7.851	2.132	2.021	5.367	1
	1.5	60	1	0	0	2.25	3.391	2.732	11.04	1
	1.5	60	2	0.891	0.588	0.717	2.882	2.629	8.818	0.9999
	1.5	60	3	1.925	1.309	1.892	2.87	2.527	8.26	0.9999
	1.5	60	4	3.428	2.743	11.238	2.902	2.507	8.249	1
	1.5	60	5	3.481	2.367	9.526	2.834	2.467	7.866	1
	3	60	1	0	0	9	5.289	3.183	15.369	1
	3	60	2	1.26	0.492	3.271	5.326	3.468	17.433	0.9999
	3	60	3	2.776	1.423	2.074	5.086	3.448	16.235	1
	3	60	4	5.212	3.651	18.225	5.06	3.532	16.717	0.9998
	3	60	5	5.189	3.028	13.962	5.006	3.465	16.03	0.9999
	6	60	1	0	0	36	7.338	3.115	11.489	1
	6	60	2	1.48	0.245	20.487	7.958	3.268	14.51	0.9998
	6	60	3	3.653	1.188	6.921	7.959	3.693	17.472	1
	6	60	4	7.78	4.236	21.11	7.949	4.045	20.16	0.9999
	6	60	5	7.401	3.26	12.59	7.967	3.961	19.56	1
Ha True	1	60	1	0	0	1	1.93	1.64	3.553	1
	1	60	2	0.63	0.506	0.393	1.555	1.31	2.024	1
	1	60	3	1.299	1	1.089	1.626	1.358	2.235	1
	1	60	4	2.162	1.704	4.253	1.663	1.395	2.387	1
	1	60	5	2.298	1.553	4.096	1.634	1.331	2.173	1
	1.5	60	1	0	0	2.25	2.626	2.039	5.422	1
	1.5	60	2	0.814	0.536	0.758	2.26	1.864	4.052	1
	1.5	60	3	1.624	1.141	1.318	2.28	1.812	3.892	1
	1.5	60	4	2.7	2.077	5.752	2.313	1.821	3.974	1
	1.5	60	5	2.84	1.868	5.284	2.262	1.739	3.604	1
	3	60	1	0	0	9	4.519	2.722	9.719	1
	3	60	2	1.224	0.472	3.377	4.559	2.949	11.126	1
	3	60	3	2.487	1.346	2.075	4.316	2.796	9.55	1
	3	60	4	4.314	2.994	10.689	4.277	2.739	9.131	1
	3	60	5	4.442	2.609	8.883	4.243	2.747	9.089	1
	6	60	1	0	0	36	6.805	2.972	9.481	1
	6	60	2	1.484	0.23	20.45	7.502	3.147	12.155	1
	6	60	3	3.503	1.241	7.777	7.418	3.44	13.843	1
	6	60	4	6.914	3.932	16.292	7.324	3.664	15.172	1
	6	60	5	6.784	3.158	10.584	7.368	3.571	14.622	1

Table C.6: Bias of $\hat{k}_T$ ($p_0 = 0.05$, $L = 1$, $TOX = 1.8$)

	true k_T	N per group	n sacr	mean and SE for estimates of *k*						
				Moon			MLE			
				mean	SD	MSE	mean	SD	MSE	prop converged
Ho True	1	60	1	0	0	1	2.346	2.172	6.531	1
	1	60	2	0.708	0.566	0.405	1.909	1.909	4.468	1
	1	60	3	1.531	1.147	1.597	1.967	1.887	4.494	0.9999
	1	60	4	2.641	2.218	7.61	2.012	1.92	4.712	1
	1	60	5	2.721	1.938	6.718	1.971	1.852	4.373	1
	1.5	60	1	0	0	2.25	3.114	2.56	9.16	1
	1.5	60	2	0.887	0.584	0.718	2.699	2.431	7.345	0.9999
	1.5	60	3	1.879	1.275	1.77	2.692	2.349	6.94	0.9999
	1.5	60	4	3.289	2.6	9.96	2.727	2.343	6.992	1
	1.5	60	5	3.344	2.25	8.464	2.669	2.308	6.691	1
	3	60	1	0	0	9	5.106	3.12	14.166	1
	3	60	2	1.258	0.491	3.274	5.174	3.395	16.253	0.9999
	3	60	3	2.761	1.413	2.055	4.954	3.376	15.215	1
	3	60	4	5.131	3.59	17.427	4.928	3.46	15.689	0.9998
	3	60	5	5.115	2.979	13.347	4.893	3.416	15.251	0.9999
	6	60	1	0	0	36	7.279	3.147	11.535	1
	6	60	2	1.48	0.245	20.488	7.901	3.299	14.499	0.9997
	6	60	3	3.648	1.189	6.945	7.896	3.697	17.257	0.9999
	6	60	4	7.745	4.22	20.856	7.882	4.042	19.88	0.9998
	6	60	5	7.375	3.253	12.472	7.912	3.973	19.438	1
Ha True	1	60	1	0	0	1	1.749	1.489	2.777	1
	1	60	2	0.627	0.499	0.388	1.46	1.197	1.645	1
	1	60	3	1.251	0.965	0.994	1.524	1.283	1.922	1
	1	60	4	2.032	1.592	3.6	1.554	1.289	1.969	1
	1	60	5	2.17	1.452	3.476	1.53	1.243	1.825	1
	1.5	60	1	0	0	2.25	2.406	1.867	4.305	1
	1.5	60	2	0.809	0.532	0.76	2.121	1.682	3.216	1
	1.5	60	3	1.58	1.113	1.244	2.148	1.669	3.205	1
	1.5	60	4	2.585	1.977	5.085	2.172	1.653	3.184	1
	1.5	60	5	2.729	1.778	4.673	2.139	1.648	3.123	1
	3	60	1	0	0	9	4.342	2.64	8.769	1
	3	60	2	1.222	0.472	3.382	4.417	2.86	10.186	0.9998
	3	60	3	2.468	1.334	2.063	4.193	2.7	8.713	1
	3	60	4	4.235	2.929	10.102	4.149	2.655	8.369	1
	3	60	5	4.365	2.554	8.382	4.12	2.664	8.351	1
	6	60	1	0	0	36	6.714	2.949	9.207	1
	6	60	2	1.484	0.23	20.451	7.419	3.148	11.923	1
	6	60	3	3.499	1.239	7.788	7.344	3.425	13.535	1
	6	60	4	6.89	3.916	16.126	7.246	3.634	14.752	0.9998
	6	60	5	6.759	3.15	10.499	7.278	3.523	14.046	1

C.3 Simulation Output: 1-sided tests

This section contains the data that are represented in the Chapter 3 Results section type I error and power Figures.

Tables C.7, C.8, C.9, and C.10 present, for varying lethality and toxicity settings, the Type I error and power comparisons among **1-sided tests** when the cumulative lifetime tumor incidence rate is $\mathbf{p_0 = 0.15}$.

Table C.11 shows the corresponding comparisons when $L = 1$, $TOX = 1.8$, and $\mathbf{p_0 = 0.05}$.

Table C.7: 1-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1$, $TOX = 1$)

	true k_T	N per Tx	n sacr	Proportion of tests with p<0.05								
				Classical tests				poly-*k*			poly-*k*-hat	
				logrank	MH	score	CA	*k*=1.5	*k*=3	*k*=6	Moon	MLE
Ho True (Type I error)	1	50	1	0.0497	0.0513	0.0498	0.0474	0.0405	0.0407	0.0406	0.0412	0.0404
	1	60	1	0.0492	0.0514	0.0526	0.0505	0.0438	0.0448	0.0445	0.0436	0.0435
	1	60	2	0.0518	0.0499	0.0515	0.0503	0.0426	0.0434	0.0424	0.0431	0.0435
	1	60	3	0.0519	0.0508	0.0514	0.0490	0.0417	0.0426	0.0423	0.0426	0.0423
	1	60	4	0.0507	0.0496	0.0502	0.0482	0.0414	0.0429	0.0415	0.0419	0.0421
	1	60	5	0.0505	0.0502	0.0510	0.0493	0.0423	0.0431	0.0429	0.0424	0.0425
	1.5	50	1	0.0515	0.0517	0.0512	0.0468	0.0402	0.0402	0.0397	0.0411	0.0406
	1.5	60	1	0.0495	0.0487	0.0510	0.0488	0.0422	0.0428	0.0426	0.0419	0.0422
	1.5	60	2	0.0496	0.0506	0.0508	0.0500	0.0427	0.0408	0.0401	0.0434	0.0429
	1.5	60	3	0.0497	0.0499	0.0505	0.0493	0.0426	0.0419	0.0423	0.0434	0.0430
	1.5	60	4	0.0494	0.0501	0.0506	0.0480	0.0416	0.0406	0.0415	0.0430	0.0422
	1.5	60	5	0.0494	0.0505	0.0496	0.0484	0.0420	0.0413	0.0420	0.0428	0.0426
	3	50	1	0.0543	0.0533	0.0513	0.0482	0.0422	0.0427	0.0421	0.0430	0.0424
	3	60	1	0.0507	0.0504	0.0511	0.0476	0.0433	0.0433	0.0445	0.0430	0.0439
	3	60	2	0.0487	0.0500	0.0486	0.0450	0.0414	0.0417	0.0411	0.0415	0.0406
	3	60	3	0.0496	0.0499	0.0492	0.0472	0.0420	0.0415	0.0413	0.0415	0.0416
	3	60	4	0.0489	0.0506	0.0489	0.0484	0.0415	0.0410	0.0417	0.0415	0.0414
	3	60	5	0.0490	0.0501	0.0498	0.0482	0.0424	0.0418	0.0413	0.0421	0.0419
	6	50	1	0.0567	0.0562	0.0540	0.0476	0.0428	0.0419	0.0427	0.0445	0.0429
	6	60	1	0.0524	0.0511	0.0512	0.0487	0.0433	0.0435	0.0442	0.0443	0.0440
	6	60	2	0.0499	0.0508	0.0499	0.0455	0.0406	0.0407	0.0416	0.0407	0.0415
	6	60	3	0.0513	0.0507	0.0497	0.0475	0.0429	0.0430	0.0433	0.0431	0.0432
	6	60	4	0.0512	0.0521	0.0503	0.0483	0.0432	0.0429	0.0428	0.0432	0.0435
	6	60	5	0.0506	0.0503	0.0494	0.0476	0.0424	0.0422	0.0418	0.0418	0.0422
Ha True (Power)	1	50	1	0.5116	0.5240	0.5368	0.5362	0.5360	0.5254	0.5182	0.5328	0.5342
	1	60	1	0.5838	0.5928	0.5916	0.5958	0.5948	0.5886	0.5844	0.5952	0.5926
	1	60	2	0.5460	0.5596	0.5650	0.5612	0.5596	0.5576	0.5540	0.5640	0.5642
	1	60	3	0.5494	0.5662	0.5714	0.5740	0.5750	0.5662	0.5592	0.5794	0.5744
	1	60	4	0.5550	0.5742	0.5776	0.5790	0.5736	0.5724	0.5646	0.5796	0.5748
	1	60	5	0.5536	0.5632	0.5740	0.5708	0.5752	0.5686	0.5590	0.5762	0.5754
	1.5	50	1	0.4954	0.5046	0.5078	0.5146	0.5118	0.5114	0.5070	0.5122	0.5122
	1.5	60	1	0.5716	0.5842	0.5774	0.5762	0.5840	0.5812	0.5758	0.5766	0.5848
	1.5	60	2	0.5230	0.5330	0.5390	0.5276	0.5396	0.5422	0.5352	0.5362	0.5382
	1.5	60	3	0.5330	0.5472	0.5528	0.5434	0.5576	0.5588	0.5474	0.5586	0.5570
	1.5	60	4	0.5350	0.5430	0.5504	0.5476	0.5628	0.5572	0.5476	0.5582	0.5566
	1.5	60	5	0.5320	0.5396	0.5548	0.5454	0.5540	0.5490	0.5460	0.5520	0.5502
	3	50	1	0.4584	0.4708	0.4828	0.4822	0.4816	0.4804	0.4748	0.4752	0.4768
	3	60	1	0.5362	0.5472	0.5500	0.5422	0.5478	0.5500	0.5470	0.5414	0.5510
	3	60	2	0.4930	0.4988	0.5092	0.4956	0.5068	0.5078	0.5048	0.5024	0.5134
	3	60	3	0.4910	0.5000	0.5074	0.5022	0.5054	0.5066	0.5084	0.5098	0.5062
	3	60	4	0.4990	0.4996	0.5150	0.5016	0.5070	0.5098	0.5102	0.5144	0.5092
	3	60	5	0.4948	0.5022	0.5092	0.5010	0.5048	0.5070	0.5076	0.5056	0.5090
	6	50	1	0.4292	0.4408	0.4460	0.4488	0.4412	0.4460	0.4478	0.4430	0.4492
	6	60	1	0.5084	0.5156	0.5174	0.4998	0.5086	0.5148	0.5124	0.5004	0.5118
	6	60	2	0.4548	0.4562	0.4664	0.4510	0.4512	0.4610	0.4636	0.4508	0.4678
	6	60	3	0.4404	0.4552	0.4630	0.4434	0.4440	0.4534	0.4572	0.4544	0.4582
	6	60	4	0.4376	0.4366	0.4562	0.4472	0.4464	0.4508	0.4530	0.4524	0.4540
	6	60	5	0.4398	0.4474	0.4586	0.4466	0.4448	0.4538	0.4576	0.4596	0.4602

Table C.8: 1-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1$, $TOX = 1.8$)

	true k_T	N per Tx	n sacr	Proportion of tests with p<0.05								
				Classical tests				poly-*k*			poly-*k*-hat	
				logrank	MH	score	CA	*k*=1.5	*k*=3	*k*=6	Moon	MLE
Ho True (Type I error)	1	50	1	0.1206	0.0499	0.0515	0.0313	0.0451	0.0566	0.0699	0.0273	0.0429
	1	60	1	0.1265	0.0516	0.0530	0.0309	0.0469	0.0607	0.0759	0.0262	0.0443
	1	60	2	0.1206	0.0484	0.0486	0.0294	0.0465	0.0596	0.0778	0.0340	0.0419
	1	60	3	0.1293	0.0487	0.0498	0.0290	0.0460	0.0609	0.0800	0.0394	0.0420
	1	60	4	0.1355	0.0494	0.0502	0.0289	0.0468	0.0624	0.0826	0.0469	0.0429
	1	60	5	0.1325	0.0497	0.0505	0.0295	0.0466	0.0611	0.0836	0.0485	0.0426
	1.5	50	1	0.1092	0.0523	0.0524	0.0258	0.0387	0.0499	0.0618	0.0231	0.0412
	1.5	60	1	0.1123	0.0487	0.0492	0.0241	0.0388	0.0514	0.0642	0.0213	0.0420
	1.5	60	2	0.1088	0.0494	0.0489	0.0248	0.0365	0.0513	0.0686	0.0295	0.0393
	1.5	60	3	0.1137	0.0503	0.0500	0.0234	0.0388	0.0531	0.0666	0.0369	0.0417
	1.5	60	4	0.1220	0.0501	0.0486	0.0232	0.0371	0.0530	0.0691	0.0423	0.0403
	1.5	60	5	0.1187	0.0509	0.0484	0.0227	0.0376	0.0523	0.0711	0.0443	0.0408
	3	50	1	0.0926	0.0568	0.0534	0.0205	0.0302	0.0396	0.0492	0.0188	0.0411
	3	60	1	0.0937	0.0518	0.0501	0.0174	0.0301	0.0396	0.0491	0.0160	0.0408
	3	60	2	0.0907	0.0537	0.0476	0.0180	0.0271	0.0372	0.0479	0.0246	0.0377
	3	60	3	0.0956	0.0522	0.0468	0.0178	0.0269	0.0367	0.0499	0.0294	0.0376
	3	60	4	0.1031	0.0522	0.0499	0.0178	0.0278	0.0386	0.0525	0.0382	0.0387
	3	60	5	0.1002	0.0534	0.0488	0.0178	0.0277	0.0385	0.0524	0.0393	0.0394
	6	50	1	0.0831	0.0601	0.0529	0.0174	0.0253	0.0318	0.0388	0.0159	0.0390
	6	60	1	0.0781	0.0539	0.0509	0.0135	0.0211	0.0284	0.0365	0.0125	0.0369
	6	60	2	0.0751	0.0558	0.0465	0.0126	0.0202	0.0268	0.0353	0.0198	0.0361
	6	60	3	0.0784	0.0521	0.0461	0.0120	0.0183	0.0255	0.0338	0.0261	0.0352
	6	60	4	0.0855	0.0546	0.0490	0.0129	0.0199	0.0262	0.0348	0.0340	0.0368
	6	60	5	0.0840	0.0572	0.0485	0.0127	0.0201	0.0261	0.0353	0.0344	0.0366
Ha True (Power)	1	50	1	0.5306	0.4968	0.5016	0.4940	0.5130	0.5206	0.5206	0.4894	0.5106
	1	60	1	0.5816	0.5504	0.5562	0.5532	0.5782	0.5820	0.5784	0.5536	0.5664
	1	60	2	0.5564	0.5326	0.5472	0.5266	0.5658	0.5656	0.5528	0.5366	0.5556
	1	60	3	0.5648	0.5442	0.5476	0.5336	0.5676	0.5716	0.5678	0.5506	0.5522
	1	60	4	0.5682	0.5434	0.5496	0.5352	0.5704	0.5766	0.5734	0.5448	0.5558
	1	60	5	0.5692	0.5378	0.5476	0.5348	0.5712	0.5742	0.5694	0.5494	0.5512
	1.5	50	1	0.4950	0.4634	0.4684	0.4516	0.4844	0.4936	0.4952	0.4512	0.4840
	1.5	60	1	0.5630	0.5372	0.5454	0.5232	0.5622	0.5610	0.5594	0.5196	0.5540
	1.5	60	2	0.5230	0.5008	0.5110	0.4896	0.5216	0.5332	0.5260	0.5096	0.5172
	1.5	60	3	0.5310	0.5048	0.5170	0.4914	0.5264	0.5300	0.5308	0.5132	0.5222
	1.5	60	4	0.5370	0.5086	0.5210	0.4982	0.5314	0.5414	0.5386	0.5194	0.5272
	1.5	60	5	0.5318	0.4998	0.5180	0.4928	0.5302	0.5372	0.5332	0.5214	0.5222
	3	50	1	0.4448	0.4150	0.4296	0.3910	0.4250	0.4384	0.4502	0.3948	0.4408
	3	60	1	0.5142	0.4958	0.5070	0.4596	0.4970	0.5124	0.5124	0.4586	0.5134
	3	60	2	0.4542	0.4368	0.4558	0.4026	0.4466	0.4610	0.4726	0.4454	0.4658
	3	60	3	0.4586	0.4512	0.4658	0.4134	0.4526	0.4706	0.4730	0.4564	0.4694
	3	60	4	0.4584	0.4468	0.4576	0.3984	0.4458	0.4630	0.4670	0.4466	0.4592
	3	60	5	0.4544	0.4376	0.4592	0.3986	0.4420	0.4560	0.4608	0.4490	0.4504
	6	50	1	0.3856	0.3634	0.3924	0.3398	0.3794	0.3988	0.4110	0.3430	0.4066
	6	60	1	0.4524	0.4410	0.4524	0.4070	0.4374	0.4524	0.4602	0.4036	0.4608
	6	60	2	0.3978	0.3814	0.4190	0.3552	0.3900	0.4122	0.4194	0.3906	0.4166
	6	60	3	0.4016	0.3926	0.4178	0.3586	0.3920	0.4048	0.4210	0.4066	0.4182
	6	60	4	0.3858	0.3644	0.3996	0.3464	0.3810	0.3928	0.4000	0.3880	0.3992
	6	60	5	0.3878	0.3658	0.3986	0.3468	0.3754	0.3912	0.4012	0.3936	0.4024

Table C.9: 1-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1.5$, $TOX = 1$)

	true k_T	N per Tx	n sacr	Proportion of tests with p<0.05								
				Classical tests				poly-*k*			poly-*k*-hat	
				logrank	MH	score	CA	*k*=1.5	*k*=3	*k*=6	Moon	MLE
Ho True (Type I error)	1	50	1	0.0497	0.0517	0.0501	0.0474	0.0405	0.0407	0.0406	0.0412	0.0401
	1	60	1	0.0498	0.0507	0.0519	0.0505	0.0438	0.0448	0.0445	0.0436	0.0433
	1	60	2	0.0526	0.0511	0.0516	0.0504	0.0427	0.0434	0.0423	0.0432	0.0434
	1	60	3	0.0514	0.0516	0.0511	0.0489	0.0417	0.0426	0.0423	0.0423	0.0423
	1	60	4	0.0505	0.0502	0.0506	0.0481	0.0416	0.0430	0.0417	0.0419	0.0415
	1	60	5	0.0505	0.0509	0.0510	0.0494	0.0425	0.0434	0.0430	0.0431	0.0424
	1.5	50	1	0.0517	0.0515	0.0503	0.0468	0.0402	0.0402	0.0397	0.0411	0.0407
	1.5	60	1	0.0497	0.0494	0.0506	0.0488	0.0422	0.0428	0.0426	0.0419	0.0423
	1.5	60	2	0.0507	0.0514	0.0509	0.0499	0.0426	0.0407	0.0400	0.0432	0.0423
	1.5	60	3	0.0502	0.0486	0.0507	0.0493	0.0426	0.0419	0.0425	0.0432	0.0428
	1.5	60	4	0.0502	0.0502	0.0504	0.0481	0.0417	0.0407	0.0417	0.0431	0.0420
	1.5	60	5	0.0496	0.0510	0.0500	0.0484	0.0421	0.0413	0.0424	0.0434	0.0427
	3	50	1	0.0539	0.0524	0.0517	0.0482	0.0422	0.0427	0.0421	0.0430	0.0422
	3	60	1	0.0508	0.0500	0.0508	0.0476	0.0433	0.0433	0.0445	0.0430	0.0438
	3	60	2	0.0492	0.0494	0.0483	0.0450	0.0413	0.0417	0.0411	0.0414	0.0407
	3	60	3	0.0498	0.0499	0.0495	0.0472	0.0420	0.0415	0.0413	0.0415	0.0414
	3	60	4	0.0495	0.0511	0.0489	0.0484	0.0415	0.0410	0.0417	0.0416	0.0413
	3	60	5	0.0494	0.0507	0.0499	0.0483	0.0424	0.0418	0.0413	0.0419	0.0416
	6	50	1	0.0566	0.0563	0.0537	0.0476	0.0428	0.0419	0.0427	0.0445	0.0429
	6	60	1	0.0521	0.0508	0.0511	0.0487	0.0433	0.0435	0.0442	0.0443	0.0440
	6	60	2	0.0502	0.0512	0.0496	0.0455	0.0406	0.0407	0.0416	0.0407	0.0417
	6	60	3	0.0511	0.0512	0.0496	0.0475	0.0429	0.0430	0.0433	0.0431	0.0435
	6	60	4	0.0512	0.0521	0.0500	0.0483	0.0433	0.0429	0.0428	0.0431	0.0433
	6	60	5	0.0507	0.0501	0.0489	0.0476	0.0424	0.0422	0.0418	0.0420	0.0423
Ha True (Power)	1	50	1	0.5144	0.5212	0.5334	0.5362	0.5360	0.5254	0.5182	0.5328	0.5308
	1	60	1	0.5826	0.5920	0.5950	0.5958	0.5948	0.5886	0.5844	0.5952	0.5946
	1	60	2	0.5474	0.5560	0.5656	0.5610	0.5594	0.5574	0.5532	0.5626	0.5636
	1	60	3	0.5566	0.5664	0.5746	0.5732	0.5732	0.5654	0.5566	0.5792	0.5738
	1	60	4	0.5574	0.5672	0.5770	0.5788	0.5738	0.5726	0.5590	0.5802	0.5762
	1	60	5	0.5558	0.5616	0.5724	0.5708	0.5754	0.5680	0.5570	0.5770	0.5788
	1.5	50	1	0.4938	0.5012	0.5116	0.5146	0.5118	0.5114	0.5070	0.5122	0.5106
	1.5	60	1	0.5722	0.5796	0.5772	0.5762	0.5840	0.5812	0.5758	0.5766	0.5820
	1.5	60	2	0.5238	0.5316	0.5356	0.5284	0.5400	0.5422	0.5352	0.5368	0.5386
	1.5	60	3	0.5316	0.5522	0.5520	0.5432	0.5572	0.5588	0.5482	0.5596	0.5544
	1.5	60	4	0.5342	0.5438	0.5522	0.5470	0.5610	0.5552	0.5482	0.5572	0.5538
	1.5	60	5	0.5316	0.5414	0.5528	0.5448	0.5536	0.5492	0.5460	0.5516	0.5486
	3	50	1	0.4602	0.4730	0.4838	0.4822	0.4816	0.4804	0.4748	0.4752	0.4748
	3	60	1	0.5356	0.5484	0.5506	0.5422	0.5478	0.5500	0.5470	0.5414	0.5512
	3	60	2	0.4924	0.5006	0.5110	0.4954	0.5068	0.5080	0.5052	0.5024	0.5136
	3	60	3	0.4914	0.4994	0.5084	0.5024	0.5056	0.5068	0.5084	0.5100	0.5094
	3	60	4	0.4934	0.4966	0.5118	0.5010	0.5056	0.5090	0.5098	0.5140	0.5096
	3	60	5	0.4940	0.5008	0.5086	0.5010	0.5046	0.5068	0.5074	0.5056	0.5088
	6	50	1	0.4296	0.4420	0.4476	0.4488	0.4412	0.4460	0.4478	0.4430	0.4490
	6	60	1	0.5100	0.5164	0.5170	0.4998	0.5086	0.5148	0.5124	0.5004	0.5100
	6	60	2	0.4534	0.4544	0.4670	0.4510	0.4512	0.4610	0.4636	0.4508	0.4680
	6	60	3	0.4400	0.4540	0.4640	0.4438	0.4440	0.4534	0.4572	0.4546	0.4582
	6	60	4	0.4372	0.4382	0.4572	0.4472	0.4462	0.4506	0.4528	0.4520	0.4536
	6	60	5	0.4402	0.4474	0.4602	0.4466	0.4450	0.4538	0.4576	0.4598	0.4580

Table C.10: 1-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$)

	true k_T	N per Tx	n sacr	Proportion of tests with p<0.05								
				Classical tests				poly-*k*			poly-*k*-hat	
				logrank	MH	score	CA	*k*=1.5	*k*=3	*k*=6	Moon	MLE
Ho True (Type I error)	1	50	1	0.1181	0.0417	0.0442	0.0313	0.0451	0.0566	0.0699	0.0273	0.0422
	1	60	1	0.1207	0.0418	0.0429	0.0309	0.0469	0.0607	0.0759	0.0262	0.0430
	1	60	2	0.1180	0.0406	0.0437	0.0293	0.0464	0.0597	0.0780	0.0338	0.0406
	1	60	3	0.1257	0.0407	0.0443	0.0290	0.0461	0.0608	0.0807	0.0380	0.0409
	1	60	4	0.1320	0.0413	0.0439	0.0286	0.0468	0.0623	0.0833	0.0452	0.0414
	1	60	5	0.1304	0.0413	0.0448	0.0295	0.0467	0.0613	0.0847	0.0473	0.0409
	1.5	50	1	0.1074	0.0469	0.0483	0.0258	0.0387	0.0499	0.0618	0.0231	0.0398
	1.5	60	1	0.1099	0.0435	0.0435	0.0241	0.0388	0.0514	0.0642	0.0213	0.0405
	1.5	60	2	0.1079	0.0443	0.0449	0.0248	0.0366	0.0513	0.0689	0.0295	0.0391
	1.5	60	3	0.1117	0.0438	0.0455	0.0233	0.0385	0.0529	0.0668	0.0366	0.0411
	1.5	60	4	0.1202	0.0443	0.0446	0.0233	0.0373	0.0530	0.0692	0.0424	0.0396
	1.5	60	5	0.1166	0.0440	0.0451	0.0227	0.0375	0.0521	0.0709	0.0433	0.0400
	3	50	1	0.0922	0.0545	0.0511	0.0205	0.0302	0.0396	0.0492	0.0188	0.0400
	3	60	1	0.0931	0.0492	0.0482	0.0174	0.0301	0.0396	0.0491	0.0160	0.0396
	3	60	2	0.0907	0.0507	0.0461	0.0180	0.0271	0.0371	0.0480	0.0247	0.0370
	3	60	3	0.0954	0.0497	0.0455	0.0178	0.0268	0.0366	0.0499	0.0293	0.0372
	3	60	4	0.1029	0.0498	0.0476	0.0178	0.0279	0.0386	0.0528	0.0378	0.0387
	3	60	5	0.1007	0.0513	0.0477	0.0178	0.0277	0.0386	0.0525	0.0385	0.0392
	6	50	1	0.0831	0.0595	0.0522	0.0174	0.0253	0.0318	0.0388	0.0159	0.0387
	6	60	1	0.0780	0.0528	0.0493	0.0135	0.0211	0.0284	0.0365	0.0125	0.0365
	6	60	2	0.0762	0.0540	0.0461	0.0126	0.0202	0.0268	0.0353	0.0198	0.0358
	6	60	3	0.0790	0.0513	0.0451	0.0120	0.0183	0.0255	0.0338	0.0261	0.0349
	6	60	4	0.0854	0.0538	0.0484	0.0129	0.0199	0.0262	0.0347	0.0341	0.0360
	6	60	5	0.0841	0.0560	0.0479	0.0127	0.0201	0.0262	0.0354	0.0344	0.0361
Ha True (Power)	1	50	1	0.5334	0.4784	0.4964	0.4940	0.5130	0.5206	0.5206	0.4894	0.5070
	1	60	1	0.5906	0.5442	0.5496	0.5532	0.5782	0.5820	0.5784	0.5536	0.5642
	1	60	2	0.5574	0.5212	0.5402	0.5266	0.5662	0.5662	0.5524	0.5378	0.5522
	1	60	3	0.5620	0.5302	0.5408	0.5334	0.5676	0.5720	0.5686	0.5506	0.5530
	1	60	4	0.5750	0.5286	0.5436	0.5350	0.5702	0.5748	0.5716	0.5474	0.5550
	1	60	5	0.5696	0.5244	0.5390	0.5342	0.5692	0.5736	0.5704	0.5512	0.5510
	1.5	50	1	0.4932	0.4568	0.4654	0.4516	0.4844	0.4936	0.4952	0.4512	0.4814
	1.5	60	1	0.5650	0.5346	0.5386	0.5232	0.5622	0.5610	0.5594	0.5196	0.5562
	1.5	60	2	0.5206	0.4964	0.5070	0.4890	0.5202	0.5322	0.5260	0.5050	0.5138
	1.5	60	3	0.5342	0.4986	0.5168	0.4912	0.5290	0.5310	0.5314	0.5104	0.5210
	1.5	60	4	0.5358	0.4994	0.5146	0.4974	0.5304	0.5408	0.5386	0.5178	0.5238
	1.5	60	5	0.5274	0.4966	0.5134	0.4920	0.5296	0.5390	0.5342	0.5126	0.5194
	3	50	1	0.4458	0.4136	0.4328	0.3910	0.4250	0.4384	0.4502	0.3948	0.4398
	3	60	1	0.5144	0.4922	0.5006	0.4596	0.4970	0.5124	0.5124	0.4586	0.5114
	3	60	2	0.4572	0.4316	0.4552	0.4028	0.4458	0.4594	0.4726	0.4458	0.4640
	3	60	3	0.4608	0.4464	0.4670	0.4130	0.4514	0.4700	0.4732	0.4568	0.4676
	3	60	4	0.4618	0.4378	0.4566	0.3984	0.4458	0.4630	0.4680	0.4440	0.4576
	3	60	5	0.4550	0.4254	0.4554	0.3984	0.4424	0.4554	0.4612	0.4488	0.4490
	6	50	1	0.3878	0.3622	0.3920	0.3398	0.3794	0.3988	0.4110	0.3430	0.4082
	6	60	1	0.4524	0.4404	0.4498	0.4070	0.4374	0.4524	0.4602	0.4036	0.4604
	6	60	2	0.3992	0.3774	0.4158	0.3552	0.3900	0.4122	0.4194	0.3906	0.4166
	6	60	3	0.4028	0.3884	0.4170	0.3586	0.3920	0.4048	0.4210	0.4062	0.4208
	6	60	4	0.3860	0.3660	0.3988	0.3462	0.3808	0.3926	0.4000	0.3878	0.3978
	6	60	5	0.3880	0.3616	0.3976	0.3468	0.3754	0.3912	0.4012	0.3928	0.4004

Table C.11: 1-sided Type I error and Power Comparisons ($p_0 = 0.05$, $L = 1$, $TOX = 1.8$)

	true k_T	N per Tx	n sacr	Proportion of tests with p<0.05								
				Classical tests				poly-*k*			poly-*k*-hat	
				logrank	MH	score	CA	*k*=1.5	*k*=3	*k*=6	Moon	MLE
Ho True (Type I error)	1	60	1	0.1328	0.0916	0.0565	0.0406	0.0499	0.0582	0.0650	0.0353	0.0512
	1	60	2	0.1399	0.1032	0.0547	0.0423	0.0506	0.0592	0.0674	0.0425	0.0498
	1	60	3	0.1402	0.0997	0.0564	0.0424	0.0505	0.0584	0.0686	0.0485	0.0507
	1	60	4	0.1456	0.0991	0.0558	0.0423	0.0510	0.0600	0.0694	0.0537	0.0506
	1	60	5	0.1448	0.1007	0.0561	0.0421	0.0508	0.0595	0.0685	0.0540	0.0500
	1.5	60	1	0.1340	0.0978	0.0552	0.0393	0.0465	0.0542	0.0605	0.0329	0.0510
	1.5	60	2	0.1435	0.1080	0.0538	0.0391	0.0448	0.0520	0.0610	0.0401	0.0469
	1.5	60	3	0.1439	0.1058	0.0557	0.0395	0.0449	0.0521	0.0614	0.0448	0.0475
	1.5	60	4	0.1479	0.1056	0.0563	0.0398	0.0454	0.0533	0.0635	0.0508	0.0476
	1.5	60	5	0.1487	0.1086	0.0563	0.0402	0.0466	0.0543	0.0639	0.0522	0.0481
	3	60	1	0.1386	0.1106	0.0542	0.0320	0.0371	0.0429	0.0491	0.0279	0.0455
	3	60	2	0.1642	0.1353	0.0535	0.0311	0.0356	0.0404	0.0488	0.0347	0.0430
	3	60	3	0.1588	0.1296	0.0554	0.0312	0.0358	0.0408	0.0490	0.0385	0.0434
	3	60	4	0.1644	0.1288	0.0554	0.0316	0.0360	0.0416	0.0499	0.0444	0.0436
	3	60	5	0.1664	0.1323	0.0560	0.0313	0.0359	0.0414	0.0494	0.0444	0.0433
	6	60	1	0.1431	0.1260	0.0566	0.0277	0.0312	0.0358	0.0407	0.0236	0.0398
	6	60	2	0.1731	0.1569	0.0582	0.0271	0.0300	0.0347	0.0403	0.0299	0.0397
	6	60	3	0.1774	0.1591	0.0571	0.0264	0.0296	0.0348	0.0394	0.0340	0.0402
	6	60	4	0.1835	0.1617	0.0574	0.0277	0.0300	0.0348	0.0395	0.0406	0.0407
	6	60	5	0.1838	0.1642	0.0566	0.0268	0.0298	0.0339	0.0391	0.0395	0.0399
Ha True (Power)	1	60	1	0.1920	0.1624	0.2376	0.2216	0.2466	0.2506	0.2534	0.2326	0.2420
	1	60	2	0.1810	0.1456	0.2290	0.2122	0.2370	0.2444	0.2480	0.2242	0.2330
	1	60	3	0.1844	0.1494	0.2268	0.2106	0.2382	0.2474	0.2510	0.2296	0.2316
	1	60	4	0.1890	0.1520	0.2320	0.2194	0.2400	0.2496	0.2508	0.2390	0.2318
	1	60	5	0.1878	0.1524	0.2304	0.2200	0.2376	0.2498	0.2498	0.2386	0.2334
	1.5	60	1	0.1790	0.1510	0.2348	0.2266	0.2324	0.2410	0.2422	0.2230	0.2268
	1.5	60	2	0.1642	0.1320	0.2262	0.2128	0.2220	0.2308	0.2364	0.2180	0.2236
	1.5	60	3	0.1674	0.1420	0.2252	0.2136	0.2244	0.2306	0.2388	0.2220	0.2244
	1.5	60	4	0.1736	0.1436	0.2250	0.2134	0.2232	0.2296	0.2364	0.2212	0.2216
	1.5	60	5	0.1698	0.1398	0.2226	0.2148	0.2242	0.2304	0.2340	0.2242	0.2228
	3	60	1	0.1480	0.1312	0.2192	0.1882	0.2160	0.2192	0.2234	0.1880	0.2192
	3	60	2	0.1294	0.1060	0.2000	0.1674	0.1982	0.2000	0.2028	0.1916	0.2030
	3	60	3	0.1338	0.1160	0.2010	0.1756	0.1980	0.2024	0.2060	0.1960	0.1992
	3	60	4	0.1408	0.1216	0.2026	0.1794	0.2000	0.2050	0.2108	0.2052	0.2026
	3	60	5	0.1332	0.1160	0.2028	0.1794	0.1982	0.2040	0.2084	0.2050	0.2022
	6	60	1	0.1266	0.1174	0.2032	0.1864	0.1942	0.2008	0.2080	0.1924	0.2090
	6	60	2	0.1072	0.1004	0.1784	0.1648	0.1692	0.1794	0.1866	0.1694	0.1862
	6	60	3	0.1128	0.0992	0.1760	0.1684	0.1666	0.1774	0.1832	0.1778	0.1824
	6	60	4	0.1154	0.1026	0.1732	0.1652	0.1638	0.1736	0.1802	0.1748	0.1766
	6	60	5	0.1120	0.0978	0.1718	0.1648	0.1624	0.1720	0.1778	0.1778	0.1776

C.4 Simulation Output: 2-sided tests

This section reports results from the same simulations as Appendix C.3 above, but with regard to 2-sided tests. These were touched on only briefly in Chapter 3; while these may be of some interest in evaluating the relative performance of the hypothesis testing methods we consider, our main focus is 1-sided tests.

Tables C.12, C.13, C.14, and C.15 present, for varying lethality and toxicity settings, the Type I error and power comparisons among **2-sided tests** when the cumulative lifetime tumor incidence rate is $\mathbf{p_0 = 0.15}$.

Table C.16 shows the corresponding comparisons when $L = 1$, $TOX = 1.8$, and $\mathbf{p_0 = 0.05}$.

Table C.12: 2-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1$, $TOX = 1$)

	true k_T	N per Tx	n sacr	Proportion of tests with p<0.05								
				Classical tests				poly-*k*			poly-*k*-hat	
				logrank	MH	score	CA	*k*=1.5	*k*=3	*k*=6	Moon	MLE
Ho True (Type I error)	1	50	1	0.0493	0.0498	0.0492	0.0469	0.0351	0.0349	0.0354	0.034	0.0353
	1	60	1	0.0498	0.0466	0.0495	0.0487	0.0365	0.0362	0.0357	0.0352	0.0364
	1	60	2	0.0449	0.0483	0.0485	0.0477	0.0358	0.0356	0.0363	0.0354	0.0355
	1	60	3	0.0467	0.0482	0.0481	0.0472	0.0356	0.0363	0.0365	0.0361	0.0361
	1	60	4	0.0462	0.0482	0.0488	0.0470	0.0358	0.0372	0.0371	0.0365	0.0362
	1	60	5	0.0460	0.0491	0.0486	0.0477	0.0359	0.0365	0.0366	0.0363	0.0360
	1.5	50	1	0.0479	0.0494	0.0495	0.0470	0.0362	0.0356	0.0353	0.0358	0.0362
	1.5	60	1	0.0509	0.0484	0.0498	0.0497	0.0383	0.0391	0.0392	0.0369	0.0386
	1.5	60	2	0.0482	0.0499	0.0496	0.0491	0.0385	0.0380	0.0390	0.0377	0.0378
	1.5	60	3	0.0497	0.0503	0.0511	0.0504	0.0381	0.0373	0.0376	0.0379	0.0378
	1.5	60	4	0.0485	0.0461	0.0496	0.0482	0.0357	0.0361	0.0353	0.0363	0.0361
	1.5	60	5	0.0483	0.0468	0.0494	0.0481	0.0360	0.0353	0.0359	0.0367	0.0365
	3	50	1	0.0466	0.0483	0.0502	0.0450	0.0340	0.0344	0.0343	0.0336	0.0336
	3	60	1	0.0501	0.0492	0.0516	0.0509	0.0368	0.0381	0.0390	0.0372	0.0382
	3	60	2	0.0470	0.0486	0.0504	0.0479	0.0370	0.0374	0.0373	0.0369	0.0377
	3	60	3	0.0469	0.0493	0.0509	0.0495	0.0377	0.0375	0.0385	0.0377	0.0382
	3	60	4	0.0475	0.0492	0.0516	0.0496	0.0373	0.0378	0.0386	0.0378	0.0378
	3	60	5	0.0477	0.0493	0.0510	0.0497	0.0379	0.0381	0.0380	0.0382	0.0383
	6	50	1	0.0463	0.0467	0.0507	0.0487	0.0357	0.0350	0.0343	0.0349	0.0344
	6	60	1	0.0497	0.0493	0.0513	0.0507	0.0365	0.0369	0.0377	0.0378	0.0374
	6	60	2	0.0492	0.0486	0.0504	0.0475	0.0389	0.0389	0.0391	0.0388	0.0390
	6	60	3	0.0487	0.0492	0.0516	0.0492	0.0379	0.0390	0.0382	0.0395	0.0385
	6	60	4	0.0473	0.0487	0.0510	0.0490	0.0363	0.0389	0.0386	0.0389	0.0384
	6	60	5	0.0493	0.0505	0.0530	0.0490	0.0381	0.0395	0.0398	0.0400	0.0402
Ha True (Power)	1	50	1	0.3926	0.4122	0.4206	0.4298	0.4196	0.4150	0.4044	0.4216	0.4204
	1	60	1	0.4562	0.4800	0.4820	0.4800	0.4846	0.4752	0.4660	0.4748	0.4826
	1	60	2	0.4424	0.4436	0.4514	0.4532	0.4538	0.4430	0.4340	0.4550	0.4518
	1	60	3	0.4424	0.4480	0.4618	0.4636	0.4564	0.4468	0.4426	0.4624	0.4562
	1	60	4	0.4426	0.4534	0.4570	0.4632	0.4592	0.4554	0.4416	0.4610	0.4604
	1	60	5	0.4426	0.4436	0.4592	0.4602	0.4582	0.4488	0.4386	0.4560	0.4542
	1.5	50	1	0.3788	0.3904	0.3982	0.4100	0.4012	0.3970	0.3892	0.3996	0.4034
	1.5	60	1	0.4360	0.4658	0.4622	0.4552	0.4650	0.4592	0.4490	0.4544	0.4592
	1.5	60	2	0.4124	0.4126	0.4222	0.4230	0.4248	0.4128	0.4028	0.4198	0.4230
	1.5	60	3	0.4108	0.4218	0.4252	0.4108	0.4278	0.4224	0.4134	0.4276	0.4242
	1.5	60	4	0.4170	0.4376	0.4370	0.4376	0.4414	0.4336	0.4268	0.4368	0.4384
	1.5	60	5	0.4142	0.4322	0.4332	0.4392	0.4336	0.4248	0.4194	0.4266	0.4322
	3	50	1	0.3702	0.3658	0.3698	0.3778	0.3720	0.3732	0.3750	0.3690	0.3766
	3	60	1	0.4254	0.4308	0.4302	0.4120	0.4264	0.4254	0.4220	0.4190	0.4260
	3	60	2	0.3812	0.3788	0.3792	0.3704	0.3776	0.3776	0.3738	0.3762	0.3790
	3	60	3	0.3854	0.3872	0.3890	0.3704	0.3864	0.3806	0.3816	0.3864	0.3832
	3	60	4	0.3810	0.3892	0.3886	0.3848	0.3846	0.3806	0.3800	0.3818	0.3810
	3	60	5	0.3828	0.3818	0.3894	0.3724	0.3834	0.3812	0.3806	0.3834	0.3822
	6	50	1	0.3432	0.3432	0.3404	0.3356	0.3402	0.3434	0.3450	0.3336	0.3458
	6	60	1	0.3956	0.3962	0.3934	0.3668	0.3932	0.3918	0.3902	0.3774	0.3902
	6	60	2	0.3332	0.3374	0.3352	0.3140	0.3352	0.3312	0.3310	0.3322	0.3304
	6	60	3	0.3348	0.3342	0.3380	0.3138	0.3314	0.3350	0.3342	0.3380	0.3342
	6	60	4	0.3362	0.3372	0.3354	0.3138	0.3278	0.3312	0.3322	0.3324	0.3318
	6	60	5	0.3266	0.3276	0.3224	0.3064	0.3216	0.3220	0.3234	0.3244	0.3226

Table C.13: 2-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1$, $TOX = 1.8$)

	true k_T	N per Tx	n sacr	Proportion of tests with p<0.05								
				Classical tests				poly-*k*			poly-*k*-hat	
				logrank	MH	score	CA	*k*=1.5	*k*=3	*k*=6	Moon	MLE
Ho True (Type I error)	1	50	1	0.0790	0.0484	0.0496	0.0518	0.0340	0.0367	0.0409	0.0363	0.0310
	1	60	1	0.0812	0.0524	0.0531	0.0562	0.0383	0.0413	0.0464	0.0409	0.0362
	1	60	2	0.0734	0.0485	0.0499	0.0546	0.0357	0.0400	0.0455	0.0358	0.0342
	1	60	3	0.0789	0.0484	0.0506	0.0554	0.0356	0.0398	0.0467	0.0365	0.0346
	1	60	4	0.0868	0.0481	0.0492	0.0553	0.0360	0.0402	0.0475	0.0368	0.0350
	1	60	5	0.0837	0.0489	0.0495	0.0550	0.0359	0.0408	0.0481	0.0361	0.0346
	1.5	50	1	0.0716	0.0496	0.0503	0.0549	0.0336	0.0359	0.0388	0.0397	0.0341
	1.5	60	1	0.0745	0.0503	0.0515	0.0624	0.0366	0.0373	0.0422	0.0464	0.0354
	1.5	60	2	0.0722	0.0483	0.0501	0.0581	0.0360	0.0372	0.0410	0.0369	0.0347
	1.5	60	3	0.0753	0.0473	0.0502	0.0594	0.0345	0.0365	0.0428	0.0354	0.0334
	1.5	60	4	0.0786	0.0468	0.0483	0.0585	0.0343	0.0359	0.0424	0.0341	0.0337
	1.5	60	5	0.0763	0.0470	0.0485	0.0592	0.0340	0.0364	0.0417	0.0348	0.0336
	3	50	1	0.0645	0.0500	0.0532	0.0638	0.0354	0.0349	0.0361	0.0462	0.0327
	3	60	1	0.0638	0.0491	0.0529	0.0712	0.0361	0.0348	0.0354	0.0526	0.0346
	3	60	2	0.0609	0.0461	0.0484	0.0671	0.0354	0.0344	0.0354	0.0379	0.0333
	3	60	3	0.0638	0.0456	0.0483	0.0683	0.0345	0.0344	0.0351	0.0349	0.0335
	3	60	4	0.0680	0.0475	0.0497	0.0680	0.0356	0.0350	0.0354	0.0340	0.0335
	3	60	5	0.0666	0.0481	0.0513	0.0682	0.0365	0.0348	0.0356	0.0352	0.0333
	6	50	1	0.0560	0.0496	0.0543	0.0719	0.0383	0.0345	0.0330	0.0537	0.0322
	6	60	1	0.0570	0.0490	0.0526	0.0790	0.0414	0.0363	0.0339	0.0588	0.0340
	6	60	2	0.0522	0.0452	0.0505	0.0705	0.0391	0.0336	0.0319	0.0389	0.0322
	6	60	3	0.0521	0.0435	0.0496	0.0759	0.0388	0.0339	0.0315	0.0336	0.0309
	6	60	4	0.0551	0.0430	0.0503	0.0761	0.0386	0.0332	0.0319	0.0316	0.0310
	6	60	5	0.0545	0.0429	0.0519	0.0764	0.0404	0.0358	0.0330	0.0331	0.0330
Ha True (Power)	1	50	1	0.5100	0.3818	0.3844	0.2842	0.4100	0.4564	0.4722	0.2888	0.4032
	1	60	1	0.5700	0.4284	0.4402	0.3274	0.4756	0.5208	0.5480	0.3234	0.4528
	1	60	2	0.5344	0.4142	0.4192	0.3080	0.4548	0.4976	0.5184	0.3806	0.4328
	1	60	3	0.5504	0.4194	0.4272	0.3108	0.4606	0.5104	0.5324	0.4142	0.4376
	1	60	4	0.5560	0.4156	0.4288	0.3200	0.4598	0.5186	0.5430	0.4482	0.4422
	1	60	5	0.5580	0.4132	0.4242	0.3186	0.4636	0.5178	0.5350	0.4556	0.4420
	1.5	50	1	0.4736	0.3510	0.3512	0.2260	0.3644	0.4118	0.4348	0.2254	0.3742
	1.5	60	1	0.5420	0.4150	0.4164	0.2538	0.4266	0.4778	0.5088	0.2532	0.4372
	1.5	60	2	0.4982	0.3792	0.3824	0.2396	0.3864	0.4472	0.4764	0.3252	0.3878
	1.5	60	3	0.5120	0.3924	0.3928	0.2398	0.3952	0.4566	0.4882	0.3642	0.4038
	1.5	60	4	0.5200	0.3988	0.3966	0.2438	0.3988	0.4590	0.4972	0.3968	0.4076
	1.5	60	5	0.5126	0.3948	0.3998	0.2424	0.3984	0.4526	0.4902	0.4104	0.4080
	3	50	1	0.4118	0.3302	0.3226	0.1442	0.2680	0.3220	0.3600	0.1466	0.3278
	3	60	1	0.4838	0.3808	0.3770	0.1522	0.3058	0.3836	0.4264	0.1554	0.3902
	3	60	2	0.4262	0.3398	0.3306	0.1254	0.2658	0.3320	0.3780	0.2406	0.3332
	3	60	3	0.4280	0.3528	0.3486	0.1294	0.2700	0.3434	0.3896	0.2998	0.3482
	3	60	4	0.4406	0.3376	0.3366	0.1368	0.2678	0.3336	0.3846	0.3172	0.3400
	3	60	5	0.4290	0.3332	0.3326	0.1268	0.2646	0.3340	0.3840	0.3252	0.3348
	6	50	1	0.3496	0.3008	0.2886	0.0876	0.1912	0.2404	0.2918	0.0884	0.2892
	6	60	1	0.4096	0.3448	0.3378	0.0956	0.2072	0.2734	0.3344	0.0930	0.3330
	6	60	2	0.3588	0.2998	0.2832	0.0744	0.1732	0.2324	0.2806	0.1724	0.2820
	6	60	3	0.3606	0.3026	0.2832	0.0694	0.1658	0.2310	0.2882	0.2332	0.2854
	6	60	4	0.3586	0.2996	0.2832	0.0714	0.1572	0.2210	0.2862	0.2606	0.2808
	6	60	5	0.3548	0.2966	0.2770	0.0702	0.1558	0.2180	0.2764	0.2596	0.2750

Table C.14: 2-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1.5$, $TOX = 1$)

	true k_T	N per Tx	n sacr	Proportion of tests with p<0.05								
				Classical tests				poly-*k*			poly-*k*-hat	
				logrank	MH	score	CA	*k*=1.5	*k*=3	*k*=6	Moon	MLE
Ho True (Type I error)	1	50	1	0.0496	0.0489	0.0507	0.0469	0.0351	0.0349	0.0354	0.0340	0.0350
	1	60	1	0.0491	0.0460	0.0489	0.0487	0.0365	0.0362	0.0357	0.0352	0.0361
	1	60	2	0.0451	0.0482	0.0488	0.0475	0.0358	0.0357	0.0364	0.0353	0.0350
	1	60	3	0.0464	0.0490	0.0485	0.0472	0.0357	0.0364	0.0368	0.0363	0.0361
	1	60	4	0.0467	0.0493	0.0483	0.0466	0.0355	0.0368	0.0370	0.0361	0.0355
	1	60	5	0.0468	0.0495	0.0482	0.0476	0.0357	0.0363	0.0364	0.0363	0.0359
	1.5	50	1	0.0483	0.0489	0.0494	0.0470	0.0362	0.0356	0.0353	0.0358	0.0358
	1.5	60	1	0.0499	0.0484	0.0497	0.0497	0.0383	0.0391	0.0392	0.0369	0.0381
	1.5	60	2	0.0486	0.0494	0.0496	0.0491	0.0385	0.0382	0.0390	0.0378	0.0382
	1.5	60	3	0.0501	0.0510	0.0510	0.0504	0.0381	0.0373	0.0377	0.0378	0.0381
	1.5	60	4	0.0486	0.0468	0.0498	0.0481	0.0355	0.0358	0.0352	0.0358	0.0358
	1.5	60	5	0.0488	0.0476	0.0494	0.0479	0.0360	0.0352	0.0360	0.0367	0.0365
	3	50	1	0.0462	0.0478	0.0492	0.0450	0.0340	0.0344	0.0343	0.0336	0.0339
	3	60	1	0.0504	0.0491	0.0521	0.0509	0.0368	0.0381	0.0390	0.0372	0.0381
	3	60	2	0.0470	0.0488	0.0511	0.0479	0.0370	0.0374	0.0374	0.0369	0.0374
	3	60	3	0.0471	0.0500	0.0511	0.0495	0.0377	0.0375	0.0386	0.0378	0.0378
	3	60	4	0.0476	0.0498	0.0513	0.0497	0.0374	0.0377	0.0385	0.0377	0.0378
	3	60	5	0.0482	0.0505	0.0508	0.0497	0.0380	0.0381	0.0381	0.0383	0.0385
	6	50	1	0.0461	0.0466	0.0511	0.0487	0.0357	0.0350	0.0343	0.0349	0.0345
	6	60	1	0.0501	0.0498	0.0514	0.0507	0.0365	0.0369	0.0377	0.0378	0.0375
	6	60	2	0.0491	0.0478	0.0503	0.0475	0.0389	0.0389	0.0391	0.0388	0.0391
	6	60	3	0.0491	0.0492	0.0519	0.0492	0.0379	0.0390	0.0382	0.0394	0.0385
	6	60	4	0.0474	0.0489	0.0510	0.0490	0.0363	0.0389	0.0386	0.0389	0.0385
	6	60	5	0.0493	0.0502	0.0532	0.0490	0.0381	0.0395	0.0398	0.0399	0.0400
Ha True (Power)	1	50	1	0.3942	0.4080	0.4174	0.4298	0.4196	0.4150	0.4044	0.4216	0.4230
	1	60	1	0.4602	0.4764	0.4780	0.4800	0.4846	0.4752	0.4660	0.4748	0.4826
	1	60	2	0.4456	0.4458	0.4530	0.4532	0.4532	0.4428	0.4332	0.4544	0.4522
	1	60	3	0.4406	0.4472	0.4634	0.4632	0.4566	0.4462	0.4416	0.4608	0.4576
	1	60	4	0.4444	0.4440	0.4606	0.4650	0.4606	0.4550	0.4418	0.4618	0.4608
	1	60	5	0.4462	0.4470	0.4598	0.4604	0.4580	0.4492	0.4396	0.4566	0.4562
	1.5	50	1	0.3792	0.3908	0.3990	0.4100	0.4012	0.3970	0.3892	0.3996	0.4024
	1.5	60	1	0.4416	0.4640	0.4628	0.4552	0.4650	0.4592	0.4490	0.4544	0.4606
	1.5	60	2	0.4118	0.4146	0.4186	0.4230	0.4242	0.4124	0.4028	0.4194	0.4234
	1.5	60	3	0.4100	0.4192	0.4240	0.4106	0.4252	0.4222	0.4130	0.4260	0.4258
	1.5	60	4	0.4190	0.4334	0.4374	0.4370	0.4412	0.4338	0.4254	0.4352	0.4388
	1.5	60	5	0.4152	0.4276	0.4346	0.4392	0.4322	0.4236	0.4208	0.4258	0.4320
	3	50	1	0.3728	0.3674	0.3726	0.3778	0.3720	0.3732	0.3750	0.3690	0.3752
	3	60	1	0.4232	0.4300	0.4294	0.4120	0.4264	0.4254	0.4220	0.4190	0.4270
	3	60	2	0.3790	0.3774	0.3782	0.3706	0.3760	0.3772	0.3734	0.3760	0.3780
	3	60	3	0.3860	0.3844	0.3882	0.3702	0.3860	0.3798	0.3814	0.3852	0.3832
	3	60	4	0.3848	0.3870	0.3880	0.3780	0.3846	0.3808	0.3802	0.3820	0.3794
	3	60	5	0.3832	0.3778	0.3910	0.3714	0.3820	0.3800	0.3790	0.3822	0.3816
	6	50	1	0.3436	0.3432	0.3392	0.3356	0.3402	0.3434	0.3450	0.3336	0.3466
	6	60	1	0.3920	0.3934	0.3928	0.3668	0.3932	0.3918	0.3902	0.3774	0.3914
	6	60	2	0.3320	0.3380	0.3354	0.3140	0.3352	0.3312	0.3310	0.3322	0.3308
	6	60	3	0.3336	0.3342	0.3356	0.3138	0.3314	0.3350	0.3344	0.3384	0.3342
	6	60	4	0.3356	0.3348	0.3342	0.3134	0.3276	0.3310	0.3318	0.3318	0.3316
	6	60	5	0.3260	0.3282	0.3214	0.3066	0.3216	0.3220	0.3230	0.3248	0.3222

Table C.15: 2-sided Type I error and Power Comparisons ($p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$)

	true k_T	N per Tx	n sacr	Proportion of tests with p<0.05								
				Classical tests				poly-*k*			poly-*k*-hat	
				logrank	MH	score	CA	*k*=1.5	*k*=3	*k*=6	Moon	MLE
Ho True (Type I error)	1	50	1	0.0760	0.0486	0.0514	0.0518	0.0340	0.0367	0.0409	0.0363	0.0317
	1	60	1	0.0794	0.0508	0.0523	0.0562	0.0383	0.0413	0.0464	0.0409	0.0359
	1	60	2	0.0723	0.0484	0.0498	0.0547	0.0355	0.0399	0.0456	0.0360	0.0335
	1	60	3	0.0774	0.0479	0.0497	0.0553	0.0355	0.0402	0.0472	0.0365	0.0345
	1	60	4	0.0849	0.0478	0.0482	0.0550	0.0361	0.0403	0.0480	0.0362	0.0334
	1	60	5	0.0828	0.0481	0.0488	0.0549	0.0358	0.0407	0.0482	0.0356	0.0341
	1.5	50	1	0.0715	0.0504	0.0523	0.0549	0.0336	0.0359	0.0388	0.0397	0.0335
	1.5	60	1	0.0727	0.0510	0.0519	0.0624	0.0366	0.0373	0.0422	0.0464	0.0354
	1.5	60	2	0.0716	0.0480	0.0503	0.0581	0.0360	0.0373	0.0411	0.0370	0.0341
	1.5	60	3	0.0743	0.0463	0.0506	0.0594	0.0345	0.0365	0.0430	0.0354	0.0340
	1.5	60	4	0.0780	0.0458	0.0478	0.0584	0.0344	0.0361	0.0427	0.0343	0.0331
	1.5	60	5	0.0768	0.0465	0.0486	0.0592	0.0341	0.0363	0.0423	0.0348	0.0332
	3	50	1	0.0642	0.0496	0.0531	0.0638	0.0354	0.0349	0.0361	0.0462	0.0331
	3	60	1	0.0635	0.0492	0.0522	0.0712	0.0361	0.0348	0.0354	0.0526	0.0340
	3	60	2	0.0603	0.0459	0.0487	0.0672	0.0354	0.0345	0.0356	0.0380	0.0332
	3	60	3	0.0633	0.0452	0.0480	0.0683	0.0344	0.0343	0.0349	0.0346	0.0329
	3	60	4	0.0682	0.0477	0.0494	0.0680	0.0356	0.0349	0.0354	0.0344	0.0334
	3	60	5	0.0674	0.0482	0.0511	0.0682	0.0365	0.0348	0.0359	0.0350	0.0338
	6	50	1	0.0558	0.0498	0.0543	0.0719	0.0383	0.0345	0.0330	0.0537	0.0320
	6	60	1	0.0570	0.0482	0.0534	0.0790	0.0414	0.0363	0.0339	0.0588	0.0342
	6	60	2	0.0518	0.0454	0.0508	0.0705	0.0391	0.0336	0.0319	0.0389	0.0324
	6	60	3	0.0521	0.0433	0.0497	0.0758	0.0388	0.0339	0.0315	0.0336	0.0311
	6	60	4	0.0553	0.0429	0.0503	0.0761	0.0386	0.0332	0.0319	0.0315	0.0311
	6	60	5	0.0541	0.0427	0.0520	0.0764	0.0404	0.0358	0.0331	0.0330	0.0332
Ha True (Power)	1	50	1	0.5130	0.3384	0.3452	0.2842	0.4100	0.4564	0.4722	0.2888	0.4018
	1	60	1	0.5766	0.3892	0.3982	0.3274	0.4756	0.5208	0.5480	0.3234	0.4476
	1	60	2	0.5396	0.3756	0.3950	0.3074	0.4550	0.4986	0.5184	0.3786	0.4246
	1	60	3	0.5496	0.3812	0.4018	0.3122	0.4612	0.5116	0.5328	0.4110	0.4338
	1	60	4	0.5662	0.3796	0.4044	0.3188	0.4602	0.5200	0.5448	0.4440	0.4354
	1	60	5	0.5598	0.3808	0.4028	0.3176	0.4652	0.5164	0.5386	0.4536	0.4372
	1.5	50	1	0.4732	0.3202	0.3244	0.2260	0.3644	0.4118	0.4348	0.2254	0.3660
	1.5	60	1	0.5416	0.3774	0.3852	0.2538	0.4266	0.4778	0.5088	0.2532	0.4296
	1.5	60	2	0.4960	0.3574	0.3662	0.2380	0.3858	0.4474	0.4758	0.3258	0.3828
	1.5	60	3	0.5092	0.3674	0.3738	0.2398	0.3938	0.4568	0.4860	0.3628	0.4008
	1.5	60	4	0.5226	0.3736	0.3818	0.2434	0.3978	0.4606	0.4990	0.3966	0.4026
	1.5	60	5	0.5074	0.3666	0.3844	0.2414	0.3970	0.4530	0.4906	0.4066	0.4030
	3	50	1	0.4138	0.3214	0.3104	0.1442	0.2680	0.3220	0.3600	0.1466	0.3262
	3	60	1	0.4808	0.3638	0.3646	0.1522	0.3058	0.3836	0.4264	0.1554	0.3884
	3	60	2	0.4298	0.3280	0.3220	0.1252	0.2656	0.3320	0.3782	0.2404	0.3314
	3	60	3	0.4312	0.3368	0.3404	0.1304	0.2696	0.3440	0.3898	0.2984	0.3432
	3	60	4	0.4444	0.3294	0.3304	0.1366	0.2676	0.3342	0.3828	0.3158	0.3382
	3	60	5	0.4326	0.3206	0.3264	0.1260	0.2634	0.3338	0.3840	0.3240	0.3324
	6	50	1	0.3482	0.2956	0.2856	0.0876	0.1912	0.2404	0.2918	0.0884	0.2882
	6	60	1	0.4108	0.3402	0.3332	0.0956	0.2072	0.2734	0.3344	0.0930	0.3304
	6	60	2	0.3600	0.2936	0.2802	0.0744	0.1732	0.2324	0.2806	0.1724	0.2800
	6	60	3	0.3602	0.2988	0.2790	0.0694	0.1658	0.2312	0.2882	0.2328	0.2844
	6	60	4	0.3586	0.2942	0.2802	0.0716	0.1572	0.2204	0.2860	0.2604	0.2798
	6	60	5	0.3550	0.2922	0.2736	0.0702	0.1560	0.2180	0.2760	0.2594	0.2736

Table C.16: 2-sided Type I error and Power Comparisons ($p_0 = 0.05$, $L = 1$, $TOX = 1.8$)

	true k_T	N per Tx	n sacr	Proportion of tests with p<0.05								
				Classical tests				poly-*k*			poly-*k*-hat	
				logrank	MH	score	CA	*k*=1.5	*k*=3	*k*=6	Moon	MLE
Ho True (Type I error)	1	60	1	0.0844	0.0589	0.0513	0.0467	0.0345	0.0354	0.0388	0.0364	0.0329
	1	60	2	0.0893	0.0635	0.0494	0.0451	0.0337	0.0361	0.0397	0.0347	0.0346
	1	60	3	0.0891	0.0628	0.0494	0.0461	0.0351	0.0369	0.0416	0.0349	0.0347
	1	60	4	0.0927	0.0632	0.0498	0.0466	0.0355	0.0370	0.0418	0.0357	0.0344
	1	60	5	0.0914	0.0630	0.0502	0.0458	0.0352	0.0367	0.0429	0.0363	0.0347
	1.5	60	1	0.0840	0.0614	0.0505	0.0462	0.0319	0.0329	0.0364	0.0345	0.0317
	1.5	60	2	0.0883	0.0672	0.0490	0.0445	0.0298	0.0330	0.0361	0.0311	0.0313
	1.5	60	3	0.0884	0.0644	0.0480	0.0442	0.0314	0.0349	0.0381	0.0314	0.0333
	1.5	60	4	0.0929	0.0656	0.0478	0.0440	0.0315	0.0341	0.0382	0.0328	0.0329
	1.5	60	5	0.0927	0.0670	0.0493	0.0442	0.0318	0.0346	0.0389	0.0340	0.0334
	3	60	1	0.0862	0.0694	0.0495	0.0467	0.0301	0.0313	0.0340	0.0356	0.0320
	3	60	2	0.0999	0.0799	0.0468	0.0439	0.0302	0.0302	0.0311	0.0308	0.0309
	3	60	3	0.0969	0.0760	0.0491	0.0456	0.0296	0.0305	0.0324	0.0300	0.0308
	3	60	4	0.0993	0.0753	0.0474	0.0459	0.0292	0.0299	0.0330	0.0299	0.0302
	3	60	5	0.0999	0.0791	0.0490	0.0468	0.0300	0.0307	0.0331	0.0312	0.0307
	6	60	1	0.0888	0.0794	0.0491	0.0473	0.0298	0.0300	0.0324	0.0360	0.0313
	6	60	2	0.1031	0.0909	0.0464	0.0434	0.0276	0.0281	0.0291	0.0280	0.0298
	6	60	3	0.1042	0.0916	0.0474	0.0440	0.0275	0.0279	0.0295	0.0276	0.0291
	6	60	4	0.1080	0.0930	0.0479	0.0436	0.0273	0.0285	0.0301	0.0289	0.0293
	6	60	5	0.1073	0.0941	0.0470	0.0425	0.0262	0.0271	0.0295	0.0290	0.0289
Ha True (Power)	1	60	1	0.1914	0.1574	0.1708	0.1392	0.1904	0.2052	0.2224	0.1406	0.1832
	1	60	2	0.1806	0.1426	0.1666	0.1348	0.1772	0.1956	0.2132	0.1550	0.1738
	1	60	3	0.1840	0.1464	0.1678	0.1362	0.1838	0.2012	0.2192	0.1694	0.1798
	1	60	4	0.1890	0.1484	0.1696	0.1378	0.1834	0.2014	0.2208	0.1868	0.1804
	1	60	5	0.1876	0.1492	0.1672	0.1370	0.1818	0.2014	0.2200	0.1888	0.1784
	1.5	60	1	0.1790	0.1466	0.1648	0.1230	0.1680	0.1864	0.2038	0.1244	0.1768
	1.5	60	2	0.1642	0.1304	0.1566	0.1206	0.1576	0.1780	0.1974	0.1394	0.1654
	1.5	60	3	0.1674	0.1390	0.1606	0.1232	0.1630	0.1814	0.2022	0.1630	0.1682
	1.5	60	4	0.1734	0.1406	0.1642	0.1246	0.1630	0.1820	0.2022	0.1728	0.1688
	1.5	60	5	0.1682	0.1382	0.1612	0.1264	0.1630	0.1814	0.2034	0.1738	0.1672
	3	60	1	0.1476	0.1290	0.1520	0.0982	0.1312	0.1486	0.1670	0.0988	0.1534
	3	60	2	0.1294	0.1058	0.1394	0.0916	0.1212	0.1374	0.1552	0.1162	0.1442
	3	60	3	0.1336	0.1152	0.1374	0.0906	0.1182	0.1378	0.1570	0.1274	0.1408
	3	60	4	0.1408	0.1216	0.1472	0.0900	0.1184	0.1412	0.1604	0.1418	0.1448
	3	60	5	0.1332	0.1154	0.1406	0.0898	0.1162	0.1380	0.1596	0.1454	0.1416
	6	60	1	0.1266	0.1174	0.1436	0.0798	0.1056	0.1234	0.1364	0.0770	0.1378
	6	60	2	0.1070	0.1000	0.1326	0.0826	0.0964	0.1112	0.1264	0.0960	0.1278
	6	60	3	0.1128	0.0990	0.1290	0.0746	0.0956	0.1132	0.1262	0.1136	0.1310
	6	60	4	0.1154	0.1026	0.1294	0.0798	0.0960	0.1104	0.1252	0.1218	0.1264
	6	60	5	0.1120	0.0978	0.1312	0.0770	0.0942	0.1108	0.1270	0.1250	0.1266

Appendix D

SUPPLEMENTAL MATERIALS FOR CHAPTER 4

D.1 Simulation Output: mean, SD, and MSE of $\widehat{k_T}$, comparing Moon and MLE estimation methods

In the tables below:

- N: sample size per treatment group
- nsacr: number of sacrifices (interim + terminal)
- *ntot*: total number of animals sacrificed at interim time points
- $MSE = \frac{1}{n_{\text{replications}}} \sum(\widehat{k_T} - k_T)^2 = Var(\widehat{k_T}) + bias^2(\widehat{k_T})$

Table D.1: Bias of $\hat{k}_T$ under Ho ($p_0 = 0.15$, $L = 1.5$, $TOX = 1$)

N per	n	n	true	Moon			MLE			true	Moon			MLE		
group	sacr	tot	k_T	mean	SD	MSE	mean	SD	MSE	k_T	mean	SD	MSE	mean	SD	MSE
50	1	0	1	0	0	1	1.592	1.143	1.656	3	0	0	9	4.138	2.324	6.696
60	1	0	1	0	0	1	1.501	0.982	1.215	3	0	0	9	3.947	2.027	5.003
60	2	10	1	0.605	0.408	0.323	1.256	0.698	0.553	3	1.279	0.378	3.106	3.777	1.923	4.302
60	2	15	1	0.668	0.424	0.291	1.201	0.616	0.420	3	1.527	0.419	2.346	3.686	1.844	3.871
60	2	20	1	0.725	0.445	0.274	1.173	0.586	0.373	3	1.747	0.465	1.787	3.591	1.722	3.313
60	3	10	1	1.044	0.718	0.517	1.308	0.747	0.653	3	2.275	1.098	1.732	3.694	1.785	3.666
60	3	16	1	1.031	0.667	0.446	1.254	0.677	0.523	3	2.405	1.098	1.559	3.549	1.587	2.821
60	3	20	1	1.038	0.659	0.435	1.231	0.656	0.483	3	2.504	1.119	1.497	3.493	1.492	2.468
60	4	12	1	1.578	1.121	1.590	1.335	0.778	0.717	3	3.477	2.067	4.500	3.681	1.720	3.423
60	4	15	1	1.520	1.033	1.338	1.309	0.748	0.654	3	3.407	1.942	3.936	3.628	1.652	3.123
60	4	21	1	1.472	0.952	1.130	1.276	0.712	0.584	3	3.353	1.798	3.356	3.543	1.537	2.656
60	5	12	1	1.694	1.028	1.539	1.319	0.758	0.676	3	3.655	1.844	3.828	3.635	1.674	3.207
60	5	16	1	1.610	0.950	1.275	1.287	0.716	0.595	3	3.570	1.711	3.250	3.561	1.570	2.780
60	5	20	1	1.571	0.903	1.142	1.266	0.695	0.554	3	3.556	1.669	3.096	3.507	1.510	2.536
70	1	0	1	0	0	1	1.445	0.886	0.983	3	0	0	9	3.774	1.737	3.615
70	2	20	1	0.687	0.411	0.267	1.166	0.547	0.327	3	1.620	0.407	2.069	3.535	1.531	2.630
70	3	20	1	0.997	0.616	0.379	1.220	0.615	0.426	3	2.396	1.022	1.409	3.449	1.341	2.001
70	4	21	1	1.403	0.895	0.963	1.261	0.669	0.515	3	3.219	1.676	2.857	3.490	1.386	2.160
70	5	20	1	1.507	0.844	0.969	1.251	0.657	0.495	3	3.431	1.553	2.596	3.455	1.378	2.105
50	1	0	1.5	0	0	2.250	2.248	1.514	2.851	6	0	0	36	6.848	3.021	9.844
60	1	0	1.5	0	0	2.250	2.123	1.303	2.085	6	0	0	36	6.578	2.729	7.783
60	2	10	1.5	0.818	0.434	0.653	1.826	0.949	1.007	6	1.568	0.203	19.682	7.177	2.865	9.589
60	2	15	1.5	0.933	0.459	0.533	1.751	0.833	0.757	6	1.930	0.234	16.622	7.414	2.976	10.855
60	2	20	1.5	1.028	0.489	0.462	1.705	0.776	0.644	6	2.278	0.270	13.929	7.629	3.088	12.187
60	3	10	1.5	1.374	0.853	0.744	1.880	1.008	1.160	6	3.423	1.112	7.875	6.980	2.882	9.265
60	3	16	1.5	1.383	0.795	0.645	1.798	0.881	0.864	6	3.903	1.239	5.932	7.053	2.947	9.796
60	3	20	1.5	1.407	0.787	0.629	1.762	0.842	0.778	6	4.219	1.335	4.954	7.078	2.957	9.908
60	4	12	1.5	2.048	1.356	2.140	1.909	1.032	1.233	6	5.925	3.147	9.906	6.825	2.805	8.546
60	4	15	1.5	1.986	1.244	1.784	1.871	0.985	1.108	6	5.961	3.111	9.680	6.863	2.798	8.574
60	4	21	1.5	1.931	1.161	1.533	1.820	0.925	0.958	6	6.103	3.128	9.796	6.870	2.796	8.573
60	5	12	1.5	2.186	1.225	1.971	1.879	0.999	1.141	6	5.996	2.724	7.420	6.955	2.904	9.343
60	5	16	1.5	2.097	1.124	1.619	1.830	0.928	0.970	6	6.076	2.772	7.691	6.924	2.864	9.058
60	5	20	1.5	2.062	1.096	1.517	1.800	0.897	0.894	6	6.226	2.841	8.123	6.955	2.867	9.129
70	1	0	1.5	0	0	2.250	2.037	1.135	1.577	6	0	0	36	6.373	2.524	6.511
70	2	20	1.5	0.968	0.444	0.480	1.697	0.738	0.583	6	2.077	0.233	15.445	7.217	2.750	9.043
70	3	20	1.5	1.348	0.735	0.563	1.748	0.792	0.689	6	3.960	1.167	5.524	6.838	2.620	7.568
70	4	21	1.5	1.854	1.088	1.309	1.799	0.862	0.833	6	5.829	2.838	8.083	6.677	2.518	6.797
70	5	20	1.5	1.993	1.021	1.285	1.784	0.834	0.777	6	5.960	2.523	6.367	6.749	2.581	7.224

Table D.2: Bias of $\hat{k}_T$ under Ho ($p_0 = 0.15$, $L = 1$, $TOX = 1.8$)

N per	n	n	true	Moon			MLE			true	Moon			MLE		
group	sacr	tot	k_T	mean	SD	MSE	mean	SD	MSE	k_T	mean	SD	MSE	mean	SD	MSE
50	1	0	1	0	0	1	1.683	1.162	1.818	3	0	0	9	4.210	2.242	6.493
60	1	0	1	0	0	1	1.574	0.972	1.274	3	0	0	9	3.996	1.970	4.874
60	2	10	1	0.568	0.387	0.337	1.321	0.703	0.597	3	1.203	0.367	3.362	3.833	1.922	4.388
60	2	15	1	0.625	0.403	0.303	1.257	0.617	0.447	3	1.432	0.405	2.623	3.723	1.833	3.883
60	2	20	1	0.680	0.426	0.284	1.219	0.586	0.391	3	1.645	0.453	2.042	3.633	1.746	3.450
60	3	10	1	1.037	0.691	0.479	1.382	0.761	0.725	3	2.209	1.052	1.734	3.742	1.753	3.625
60	3	16	1	1.028	0.654	0.429	1.320	0.676	0.559	3	2.343	1.068	1.573	3.610	1.564	2.819
60	3	20	1	1.042	0.649	0.423	1.295	0.658	0.520	3	2.442	1.098	1.517	3.550	1.503	2.562
60	4	12	1	1.615	1.083	1.552	1.408	0.778	0.773	3	3.435	1.949	3.986	3.738	1.673	3.343
60	4	15	1	1.566	1.013	1.347	1.382	0.749	0.708	3	3.386	1.863	3.618	3.694	1.638	3.165
60	4	21	1	1.521	0.958	1.190	1.350	0.714	0.633	3	3.338	1.755	3.194	3.610	1.536	2.731
60	5	12	1	1.758	1.003	1.582	1.395	0.761	0.735	3	3.637	1.734	3.412	3.694	1.628	3.130
60	5	16	1	1.680	0.941	1.348	1.360	0.718	0.645	3	3.588	1.689	3.197	3.636	1.595	2.948
60	5	20	1	1.641	0.903	1.227	1.337	0.696	0.598	3	3.561	1.635	2.989	3.574	1.523	2.648
70	1	0	1	0	0	1	1.511	0.855	0.993	3	0	0	9	3.827	1.670	3.473
70	2	20	1	0.644	0.391	0.280	1.220	0.545	0.346	3	1.524	0.393	2.334	3.566	1.488	2.535
70	3	20	1	0.992	0.604	0.365	1.285	0.615	0.459	3	2.332	1.003	1.453	3.505	1.327	2.017
70	4	21	1	1.457	0.897	1.014	1.330	0.667	0.554	3	3.214	1.651	2.770	3.545	1.365	2.161
70	5	20	1	1.580	0.843	1.048	1.321	0.653	0.529	3	3.431	1.503	2.445	3.515	1.345	2.075
50	1	0	1.5	0	0	2.250	2.339	1.475	2.877	6	0	0	36	7.034	3.032	10.261
60	1	0	1.5	0	0	2.250	2.198	1.256	2.064	6	0	0	36	6.679	2.710	7.805
60	2	10	1.5	0.766	0.412	0.709	1.886	0.928	1.010	6	1.486	0.194	20.412	7.234	2.907	9.973
60	2	15	1.5	0.870	0.439	0.590	1.803	0.818	0.760	6	1.823	0.221	17.498	7.442	2.980	10.960
60	2	20	1.5	0.962	0.470	0.511	1.749	0.760	0.640	6	2.159	0.258	14.817	7.678	3.114	12.508
60	3	10	1.5	1.347	0.807	0.675	1.944	0.974	1.145	6	3.335	1.081	8.271	7.015	2.847	9.134
60	3	16	1.5	1.362	0.775	0.619	1.860	0.861	0.871	6	3.803	1.210	6.289	7.089	2.906	9.630
60	3	20	1.5	1.389	0.776	0.615	1.826	0.838	0.809	6	4.117	1.314	5.271	7.137	2.964	10.078
60	4	12	1.5	2.066	1.301	2.012	1.976	0.998	1.224	6	5.825	2.980	8.911	6.903	2.800	8.656
60	4	15	1.5	2.013	1.225	1.763	1.939	0.952	1.099	6	5.872	3.005	9.046	6.915	2.785	8.596
60	4	21	1.5	1.965	1.152	1.543	1.895	0.905	0.975	6	5.993	2.991	8.945	6.917	2.754	8.426
60	5	12	1.5	2.234	1.194	1.964	1.949	0.966	1.135	6	5.901	2.578	6.655	6.965	2.821	8.888
60	5	16	1.5	2.154	1.124	1.692	1.905	0.909	0.990	6	6.016	2.639	6.966	6.990	2.834	9.009
60	5	20	1.5	2.115	1.085	1.555	1.872	0.884	0.919	6	6.139	2.708	7.354	6.983	2.844	9.054
70	1	0	1.5	0	0	2.250	2.108	1.082	1.540	6	0	0	36	6.401	2.485	6.337
70	2	20	1.5	0.907	0.426	0.534	1.749	0.716	0.574	6	1.963	0.221	16.346	7.256	2.787	9.342
70	3	20	1.5	1.330	0.722	0.550	1.812	0.778	0.703	6	3.866	1.153	5.884	6.872	2.619	7.620
70	4	21	1.5	1.885	1.085	1.325	1.869	0.843	0.847	6	5.748	2.784	7.814	6.701	2.489	6.686
70	5	20	1.5	2.042	1.005	1.304	1.852	0.815	0.788	6	5.872	2.443	5.982	6.771	2.529	6.988

Table D.3: Bias of $\hat{k}_T$ under Ho ($p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$)

N per	n	n	true	Moon			MLE			true	Moon			MLE		
group	sacr	tot	k_T	mean	SD	MSE	mean	SD	MSE	k_T	mean	SD	MSE	mean	SD	MSE
50	1	0	1	0	0	1	1.522	1.008	1.288	3	0	0	9	4.035	2.133	5.619
60	1	0	1	0	0	1	1.429	0.877	0.954	3	0	0	9	3.827	1.878	4.211
60	2	10	1	0.564	0.386	0.339	1.244	0.676	0.516	3	1.203	0.365	3.361	3.719	1.817	3.816
60	2	15	1	0.624	0.4	0.302	1.196	0.603	0.402	3	1.432	0.404	2.62	3.636	1.74	3.431
60	2	20	1	0.679	0.424	0.283	1.169	0.576	0.361	3	1.645	0.452	2.04	3.564	1.667	3.096
60	3	10	1	0.995	0.677	0.458	1.289	0.721	0.604	3	2.193	1.049	1.751	3.642	1.687	3.257
60	3	16	1	0.998	0.644	0.415	1.244	0.657	0.491	3	2.334	1.061	1.57	3.539	1.521	2.604
60	3	20	1	1.014	0.638	0.408	1.228	0.643	0.465	3	2.433	1.093	1.517	3.488	1.463	2.38
60	4	12	1	1.511	1.028	1.319	1.31	0.737	0.639	3	3.367	1.921	3.825	3.637	1.621	3.034
60	4	15	1	1.474	0.963	1.152	1.29	0.713	0.593	3	3.327	1.836	3.479	3.6	1.585	2.873
60	4	21	1	1.456	0.923	1.06	1.269	0.693	0.553	3	3.304	1.742	3.125	3.539	1.502	2.548
60	5	12	1	1.642	0.959	1.332	1.3	0.724	0.615	3	3.563	1.711	3.245	3.601	1.589	2.884
60	5	16	1	1.581	0.896	1.141	1.275	0.692	0.554	3	3.523	1.658	3.023	3.553	1.533	2.657
60	5	20	1	1.557	0.861	1.052	1.259	0.677	0.526	3	3.51	1.62	2.884	3.506	1.485	2.461
70	1	0	1	0	0	1	1.377	0.787	0.761	3	0	0	9	3.677	1.602	3.023
70	2	20	1	0.643	0.389	0.279	1.165	0.539	0.318	3	1.524	0.393	2.334	3.504	1.432	2.304
70	3	20	1	0.963	0.595	0.355	1.214	0.605	0.412	3	2.326	1.001	1.455	3.443	1.3	1.888
70	4	21	1	1.378	0.859	0.882	1.249	0.65	0.484	3	3.172	1.639	2.715	3.478	1.344	2.033
70	5	20	1	1.485	0.805	0.884	1.242	0.639	0.467	3	3.376	1.489	2.357	3.444	1.32	1.939
50	1	0	1.5	0	0	2.25	2.144	1.325	2.17	6	0	0	36	6.897	2.991	9.75
60	1	0	1.5	0	0	2.25	2.027	1.137	1.569	6	0	0	36	6.55	2.659	7.372
60	2	10	1.5	0.764	0.412	0.711	1.807	0.891	0.888	6	1.486	0.193	20.413	7.13	2.863	9.471
60	2	15	1.5	0.87	0.436	0.587	1.744	0.808	0.713	6	1.823	0.221	17.498	7.338	2.932	10.386
60	2	20	1.5	0.962	0.468	0.508	1.702	0.75	0.603	6	2.159	0.257	14.817	7.579	3.062	11.87
60	3	10	1.5	1.317	0.804	0.679	1.85	0.933	0.992	6	3.331	1.08	8.291	6.918	2.789	8.62
60	3	16	1.5	1.34	0.766	0.613	1.787	0.846	0.798	6	3.801	1.207	6.293	7.015	2.867	9.248
60	3	20	1.5	1.371	0.767	0.606	1.763	0.824	0.749	6	4.116	1.309	5.264	7.071	2.914	9.64
60	4	12	1.5	1.97	1.253	1.791	1.874	0.957	1.055	6	5.799	2.973	8.878	6.829	2.771	8.365
60	4	15	1.5	1.931	1.184	1.587	1.846	0.922	0.969	6	5.844	2.981	8.912	6.84	2.744	8.236
60	4	21	1.5	1.907	1.12	1.419	1.815	0.887	0.887	6	5.977	2.975	8.851	6.868	2.721	8.155
60	5	12	1.5	2.129	1.154	1.727	1.854	0.931	0.992	6	5.874	2.569	6.614	6.887	2.795	8.599
60	5	16	1.5	2.064	1.09	1.508	1.82	0.887	0.89	6	5.986	2.619	6.861	6.909	2.77	8.498
60	5	20	1.5	2.039	1.051	1.396	1.795	0.865	0.835	6	6.117	2.692	7.262	6.932	2.818	8.81
70	1	0	1.5	0	0	2.25	1.95	0.998	1.198	6	0	0	36	6.28	2.441	6.038
70	2	20	1.5	0.905	0.424	0.534	1.694	0.703	0.531	6	1.963	0.221	16.346	7.173	2.746	8.916
70	3	20	1.5	1.308	0.712	0.544	1.738	0.758	0.632	6	3.863	1.151	5.889	6.815	2.583	7.334
70	4	21	1.5	1.816	1.049	1.2	1.781	0.818	0.748	6	5.73	2.77	7.746	6.656	2.474	6.551
70	5	20	1.5	1.956	0.97	1.149	1.771	0.797	0.708	6	5.85	2.434	5.944	6.712	2.504	6.778

Table D.4: Bias of $\hat{k}_T$ under Ha ($p_0 = 0.15$, $L = 1.5$, $TOX = 1$)

N per	n	n	true	Moon			MLE			true	Moon			MLE		
group	sacr	tot	k_T	mean	SD	MSE	mean	SD	MSE	k_T	mean	SD	MSE	mean	SD	MSE
50	1	0	1	0	0	1	1.415	0.912	1.004	3	0	0	9	3.817	1.822	3.986
60	1	0	1	0	0	1	1.348	0.793	0.749	3	0	0	9	3.670	1.604	3.022
60	2	10	1	0.596	0.362	0.294	1.169	0.552	0.333	3	1.284	0.328	3.053	3.482	1.386	2.152
60	2	15	1	0.665	0.375	0.253	1.129	0.504	0.270	3	1.536	0.360	2.271	3.423	1.293	1.850
60	2	20	1	0.726	0.391	0.228	1.112	0.479	0.242	3	1.761	0.398	1.693	3.374	1.241	1.680
60	3	10	1	0.959	0.624	0.390	1.211	0.607	0.412	3	2.185	0.952	1.570	3.447	1.313	1.923
60	3	16	1	0.973	0.576	0.332	1.174	0.556	0.339	3	2.344	0.927	1.290	3.352	1.149	1.443
60	3	20	1	0.986	0.565	0.320	1.153	0.536	0.310	3	2.453	0.943	1.188	3.311	1.099	1.304
60	4	12	1	1.378	0.911	0.973	1.229	0.627	0.446	3	3.159	1.652	2.753	3.455	1.295	1.885
60	4	15	1	1.342	0.850	0.840	1.211	0.610	0.416	3	3.115	1.531	2.356	3.417	1.242	1.715
60	4	21	1	1.341	0.801	0.758	1.191	0.584	0.377	3	3.117	1.445	2.101	3.353	1.161	1.471
60	5	12	1	1.491	0.837	0.942	1.219	0.618	0.430	3	3.369	1.484	2.337	3.421	1.277	1.806
60	5	16	1	1.446	0.784	0.813	1.195	0.589	0.385	3	3.328	1.381	2.015	3.368	1.202	1.580
60	5	20	1	1.425	0.748	0.739	1.181	0.574	0.363	3	3.340	1.353	1.946	3.333	1.140	1.409
70	1	0	1	0	0	1	1.301	0.710	0.595	3	0	0	9	3.555	1.428	2.347
70	2	20	1	0.679	0.354	0.228	1.093	0.449	0.210	3	1.622	0.350	2.021	3.308	1.100	1.305
70	3	20	1	0.920	0.522	0.279	1.131	0.504	0.271	3	2.331	0.857	1.182	3.275	1.015	1.106
70	4	21	1	1.252	0.737	0.606	1.158	0.547	0.324	3	2.999	1.362	1.854	3.323	1.091	1.296
70	5	20	1	1.342	0.689	0.591	1.153	0.536	0.310	3	3.201	1.233	1.560	3.295	1.068	1.226
50	1	0	1.5	0	0	2.250	2.013	1.151	1.588	6	0	0	36	6.776	2.697	7.877
60	1	0	1.5	0	0	2.250	1.927	0.994	1.171	6	0	0	36	6.504	2.375	5.893
60	2	10	1.5	0.810	0.378	0.619	1.696	0.705	0.536	6	1.576	0.188	19.604	6.970	2.501	7.193
60	2	15	1.5	0.928	0.397	0.484	1.647	0.627	0.414	6	1.938	0.216	16.548	7.122	2.625	8.150
60	2	20	1.5	1.031	0.418	0.394	1.622	0.588	0.361	6	2.286	0.248	13.853	7.287	2.743	9.177
60	3	10	1.5	1.280	0.730	0.580	1.741	0.760	0.635	6	3.378	0.991	7.858	6.748	2.437	6.498
60	3	16	1.5	1.320	0.683	0.498	1.691	0.689	0.512	6	3.871	1.072	5.682	6.718	2.300	5.803
60	3	20	1.5	1.351	0.679	0.484	1.663	0.660	0.462	6	4.189	1.147	4.595	6.749	2.341	6.038
60	4	12	1.5	1.819	1.087	1.283	1.764	0.781	0.680	6	5.468	2.578	6.929	6.626	2.322	5.782
60	4	15	1.5	1.789	1.028	1.140	1.740	0.761	0.638	6	5.557	2.522	6.558	6.615	2.280	5.577
60	4	21	1.5	1.776	0.956	0.991	1.708	0.727	0.572	6	5.750	2.456	6.094	6.588	2.190	5.142
60	5	12	1.5	1.971	1.009	1.239	1.751	0.775	0.664	6	5.614	2.295	5.417	6.658	2.355	5.976
60	5	16	1.5	1.915	0.942	1.060	1.719	0.734	0.587	6	5.749	2.264	5.188	6.671	2.300	5.738
60	5	20	1.5	1.900	0.896	0.963	1.698	0.709	0.542	6	5.913	2.250	5.068	6.653	2.219	5.348
70	1	0	1.5	0	0	2.250	1.868	0.893	0.933	6	0	0	36	6.314	2.203	4.950
70	2	20	1.5	0.969	0.375	0.423	1.611	0.562	0.328	6	2.088	0.211	15.350	7.002	2.421	6.864
70	3	20	1.5	1.283	0.624	0.436	1.648	0.622	0.409	6	3.917	0.990	5.319	6.562	2.110	4.767
70	4	21	1.5	1.685	0.884	0.815	1.683	0.682	0.498	6	5.484	2.269	5.412	6.442	2.010	4.233
70	5	20	1.5	1.817	0.833	0.794	1.677	0.670	0.480	6	5.645	2.062	4.375	6.480	2.020	4.310

Table D.5: Bias of $\hat{k}_T$ under Ha ($p_0 = 0.15$, $L = 1$, $TOX = 1.8$)

N per	n	n	true	Moon			MLE			true	Moon			MLE		
group	sacr	tot	k_T	mean	SD	MSE	mean	SD	MSE	k_T	mean	SD	MSE	mean	SD	MSE
50	1	0	1	0	0	1	1.323	0.801	0.746	3	0	0	9	3.670	1.701	3.343
60	1	0	1	0	0	1	1.247	0.689	0.536	3	0	0	9	3.464	1.414	2.215
60	2	10	1	0.544	0.346	0.328	1.125	0.511	0.277	3	1.200	0.323	3.345	3.368	1.284	1.784
60	2	15	1	0.602	0.360	0.288	1.095	0.479	0.238	3	1.430	0.355	2.592	3.320	1.223	1.597
60	2	20	1	0.654	0.377	0.262	1.073	0.462	0.219	3	1.641	0.398	2.005	3.268	1.163	1.425
60	3	10	1	0.911	0.588	0.354	1.154	0.561	0.338	3	2.064	0.920	1.722	3.316	1.215	1.576
60	3	16	1	0.921	0.557	0.317	1.123	0.520	0.286	3	2.209	0.915	1.463	3.241	1.104	1.276
60	3	20	1	0.925	0.552	0.310	1.105	0.507	0.268	3	2.298	0.923	1.345	3.198	1.059	1.161
60	4	12	1	1.331	0.855	0.840	1.163	0.571	0.352	3	2.972	1.518	2.303	3.305	1.199	1.531
60	4	15	1	1.307	0.810	0.750	1.152	0.558	0.334	3	2.949	1.472	2.169	3.279	1.165	1.434
60	4	21	1	1.280	0.761	0.657	1.132	0.541	0.310	3	2.916	1.368	1.877	3.223	1.101	1.261
60	5	12	1	1.479	0.802	0.873	1.156	0.563	0.341	3	3.226	1.387	1.975	3.290	1.173	1.460
60	5	16	1	1.430	0.769	0.777	1.141	0.545	0.317	3	3.196	1.333	1.815	3.241	1.122	1.316
60	5	20	1	1.395	0.736	0.698	1.127	0.537	0.305	3	3.179	1.297	1.714	3.208	1.094	1.240
70	1	0	1	0	0	1	1.210	0.614	0.421	3	0	0	9	3.387	1.256	1.727
70	2	20	1	0.611	0.338	0.266	1.066	0.433	0.191	3	1.510	0.346	2.340	3.228	1.052	1.159
70	3	20	1	0.866	0.501	0.269	1.094	0.480	0.239	3	2.193	0.846	1.367	3.185	0.982	0.998
70	4	21	1	1.198	0.694	0.520	1.114	0.512	0.276	3	2.818	1.279	1.669	3.207	1.028	1.099
70	5	20	1	1.320	0.672	0.554	1.110	0.509	0.271	3	3.059	1.185	1.407	3.182	1.009	1.051
50	1	0	1.5	0	0	2.250	1.885	1.001	1.151	6	0	0	36	6.571	2.646	7.326
60	1	0	1.5	0	0	2.250	1.799	0.873	0.852	6	0	0	36	6.262	2.310	5.403
60	2	10	1.5	0.744	0.367	0.707	1.647	0.685	0.491	6	1.494	0.179	20.333	6.734	2.439	6.485
60	2	15	1.5	0.848	0.386	0.574	1.603	0.606	0.378	6	1.829	0.207	17.438	6.968	2.638	7.898
60	2	20	1.5	0.940	0.410	0.482	1.573	0.575	0.336	6	2.165	0.240	14.765	7.084	2.694	8.431
60	3	10	1.5	1.206	0.690	0.562	1.671	0.717	0.543	6	3.260	0.976	8.461	6.484	2.300	5.521
60	3	16	1.5	1.239	0.659	0.503	1.625	0.653	0.442	6	3.722	1.070	6.332	6.559	2.355	5.855
60	3	20	1.5	1.258	0.658	0.491	1.602	0.638	0.418	6	4.014	1.160	5.291	6.581	2.359	5.902
60	4	12	1.5	1.738	1.023	1.102	1.683	0.728	0.564	6	5.224	2.394	6.333	6.368	2.185	4.907
60	4	15	1.5	1.707	0.972	0.986	1.662	0.704	0.521	6	5.298	2.400	6.254	6.378	2.200	4.980
60	4	21	1.5	1.673	0.912	0.862	1.635	0.682	0.483	6	5.426	2.377	5.981	6.378	2.134	4.695
60	5	12	1.5	1.914	0.953	1.078	1.673	0.726	0.556	6	5.401	2.147	4.966	6.431	2.233	5.169
60	5	16	1.5	1.861	0.900	0.940	1.645	0.682	0.485	6	5.565	2.183	4.952	6.469	2.252	5.289
60	5	20	1.5	1.826	0.869	0.861	1.626	0.669	0.464	6	5.692	2.230	5.067	6.480	2.240	5.248
70	1	0	1.5	0	0	2.250	1.757	0.793	0.694	6	0	0	36	6.129	2.087	4.373
70	2	20	1.5	0.881	0.364	0.516	1.570	0.534	0.290	6	1.971	0.202	16.276	6.771	2.372	6.219
70	3	20	1.5	1.194	0.603	0.458	1.594	0.593	0.360	6	3.767	1.009	6.003	6.363	2.039	4.290
70	4	21	1.5	1.594	0.831	0.700	1.621	0.642	0.427	6	5.233	2.237	5.592	6.205	1.923	3.738
70	5	20	1.5	1.752	0.796	0.697	1.612	0.627	0.406	6	5.445	2.020	4.387	6.263	1.937	3.822

Table D.6: Bias of $\hat{k}_T$ under Ha ($p_0 = 0.15$, $L = 1.5$, $TOX = 1.8$)

N per	n	n	true	Moon			MLE			true	Moon			MLE		
group	sacr	tot	k_T	mean	SD	MSE	mean	SD	MSE	k_T	mean	SD	MSE	mean	SD	MSE
50	1	0	1	0	0	1	1.192	0.754	0.606	3	0	0	9	3.505	1.641	2.946
60	1	0	1	0	0	1	1.135	0.666	0.461	3	0	0	9	3.322	1.374	1.991
60	2	10	1	0.543	0.346	0.329	1.058	0.517	0.271	3	1.2	0.323	3.344	3.288	1.236	1.61
60	2	15	1	0.603	0.359	0.287	1.036	0.482	0.233	3	1.43	0.354	2.59	3.259	1.177	1.453
60	2	20	1	0.655	0.374	0.259	1.023	0.464	0.216	3	1.641	0.396	2.003	3.224	1.13	1.327
60	3	10	1	0.886	0.58	0.35	1.071	0.558	0.316	3	2.055	0.914	1.728	3.235	1.189	1.468
60	3	16	1	0.898	0.547	0.31	1.051	0.52	0.273	3	2.202	0.91	1.464	3.185	1.1	1.244
60	3	20	1	0.903	0.54	0.301	1.039	0.507	0.259	3	2.294	0.919	1.342	3.149	1.054	1.132
60	4	12	1	1.26	0.826	0.749	1.077	0.569	0.329	3	2.927	1.493	2.234	3.225	1.176	1.432
60	4	15	1	1.245	0.783	0.672	1.07	0.557	0.315	3	2.908	1.447	2.101	3.207	1.151	1.367
60	4	21	1	1.222	0.734	0.588	1.058	0.539	0.294	3	2.892	1.351	1.836	3.164	1.086	1.205
60	5	12	1	1.386	0.773	0.746	1.074	0.563	0.322	3	3.172	1.372	1.911	3.212	1.151	1.37
60	5	16	1	1.349	0.734	0.661	1.064	0.546	0.302	3	3.144	1.317	1.753	3.178	1.107	1.256
60	5	20	1	1.321	0.703	0.598	1.052	0.535	0.289	3	3.14	1.285	1.67	3.151	1.081	1.192
70	1	0	1	0	0	1	1.097	0.605	0.375	3	0	0	9	3.238	1.209	1.519
70	2	20	1	0.61	0.336	0.266	1.009	0.433	0.188	3	1.51	0.345	2.339	3.173	1.013	1.055
70	3	20	1	0.846	0.495	0.268	1.022	0.481	0.232	3	2.188	0.842	1.368	3.127	0.977	0.97
70	4	21	1	1.14	0.675	0.476	1.032	0.507	0.258	3	2.787	1.269	1.656	3.136	1.015	1.048
70	5	20	1	1.244	0.651	0.483	1.03	0.506	0.257	3	3.011	1.174	1.379	3.116	0.994	1.002
50	1	0	1.5	0	0	2.25	1.736	0.929	0.918	6	0	0	36	6.438	2.594	6.92
60	1	0	1.5	0	0	2.25	1.674	0.852	0.756	6	0	0	36	6.124	2.261	5.125
60	2	10	1.5	0.743	0.366	0.707	1.573	0.67	0.454	6	1.494	0.179	20.335	6.628	2.407	6.189
60	2	15	1.5	0.849	0.385	0.572	1.545	0.604	0.366	6	1.829	0.207	17.439	6.873	2.606	7.552
60	2	20	1.5	0.94	0.409	0.481	1.526	0.575	0.331	6	2.165	0.24	14.764	6.994	2.647	7.994
60	3	10	1.5	1.183	0.683	0.566	1.585	0.71	0.511	6	3.258	0.975	8.471	6.398	2.282	5.364
60	3	16	1.5	1.219	0.652	0.504	1.554	0.65	0.425	6	3.722	1.066	6.327	6.502	2.337	5.713
60	3	20	1.5	1.244	0.65	0.488	1.538	0.635	0.405	6	4.014	1.157	5.283	6.526	2.327	5.689
60	4	12	1.5	1.672	0.998	1.024	1.591	0.719	0.524	6	5.207	2.389	6.334	6.293	2.171	4.798
60	4	15	1.5	1.641	0.936	0.896	1.576	0.696	0.49	6	5.281	2.397	6.262	6.308	2.176	4.828
60	4	21	1.5	1.622	0.89	0.807	1.561	0.679	0.464	6	5.414	2.368	5.95	6.322	2.113	4.567
60	5	12	1.5	1.827	0.934	0.98	1.584	0.718	0.522	6	5.376	2.14	4.969	6.356	2.218	5.046
60	5	16	1.5	1.78	0.872	0.839	1.565	0.676	0.461	6	5.544	2.181	4.962	6.411	2.247	5.216
60	5	20	1.5	1.757	0.842	0.774	1.551	0.664	0.444	6	5.674	2.222	5.041	6.423	2.213	5.076
70	1	0	1.5	0	0	2.25	1.621	0.766	0.601	6	0	0	36	5.994	2.042	4.17
70	2	20	1.5	0.881	0.362	0.514	1.516	0.534	0.286	6	1.971	0.202	16.275	6.689	2.325	5.881
70	3	20	1.5	1.179	0.6	0.463	1.526	0.595	0.355	6	3.767	1.007	6.001	6.311	2.01	4.136
70	4	21	1.5	1.54	0.812	0.661	1.541	0.639	0.41	6	5.224	2.23	5.575	6.159	1.916	3.697
70	5	20	1.5	1.679	0.771	0.627	1.534	0.628	0.395	6	5.423	2.015	4.391	6.211	1.924	3.744

VITA

Anna Korpak is a biostatistician living in Seattle, WA at the time of this publication. She welcomes your comments to akorpak@uw.edu.