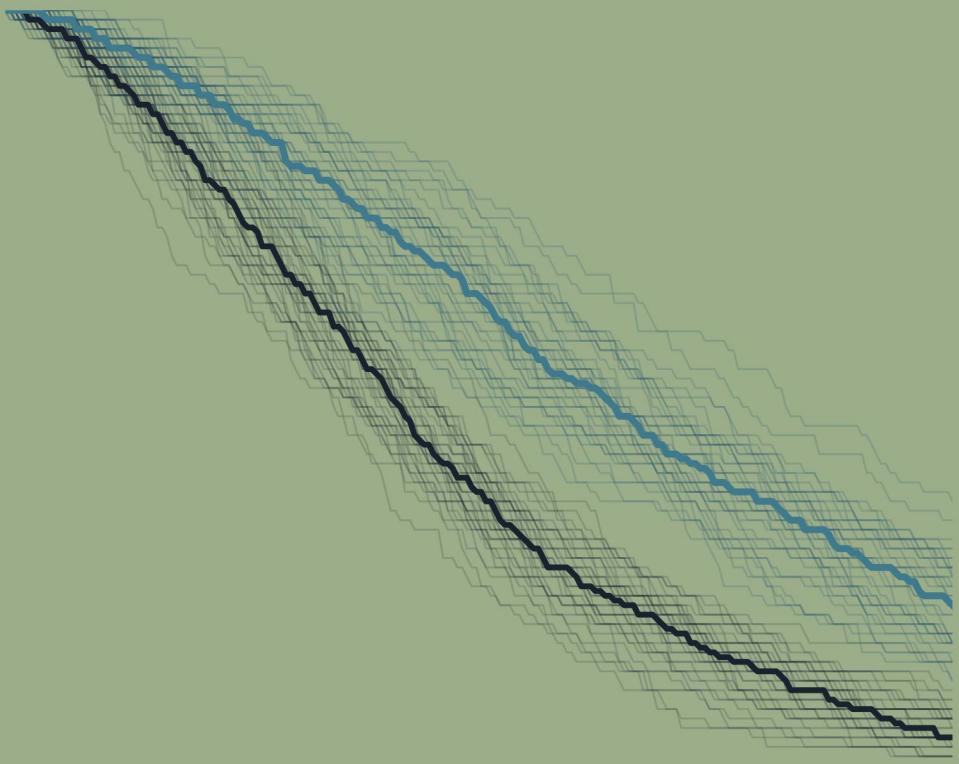
APPLIED MONTE CARLO 1



Monte Carlo for Clinical Trial Biostatisticians

Estimands, Adaptive Designs, Bayesian Borrowing, and External Controls in R

Ingrid Voss

ODIN PRESS

Monte Carlo for Clinical Trial Biostatisticians

Paperback version, First edition: 2026.
ISBN: forthcoming (Bowker allocation pending)
Trim Size: 6.14×9.21 inch.

Published by Odin Press.
`https://odinpress.org`
This book was typeset in LaTeX by Odin Press from a manuscript submitted by the author.

Printed and Bound in the United States of America.

9 8 7 6 5 4 3 2 1 0

Preface

When a clinical-trial protocol says the trial is "powered at 0.90," it is making a quiet promise: that if the treatment genuinely works in the way the sponsor expects, the trial will detect it with probability 0.90. The promise is plural rather than singular. It rests on a specific story about how the patients in the trial behave (the data-generating process), a specific test that decides whether the trial wins or loses (the test statistic), a specific rule for what counts as winning (the decision rule), and a specific computer program that produces the headline number itself (the simulator). Change any one of those four, and the 0.90 changes by enough to matter. The job of the modern biostatistician is to make every word of that promise operational: to choose the data-generating process against the realistic clinical evidence, to choose the test statistic against the regulator's expectations, to choose the decision rule against the trial's operational constraints, and to write a simulator whose output is reproducible to the digit by a regulator running the same code on a clean machine five years after the trial's design lock. This book is about that discipline.

The discipline has matured rapidly. When I started in biostatistics in the early 1990s, group-sequential boundaries were drawn by hand on spending-function plots, the proportional-hazards assumption was the default for survival endpoints regardless of the underlying biology, and "Monte Carlo simulation" was a research-statistics activity that rarely surfaced in regulatory briefings. By the time I retired in early 2024, the simulator had become the operative tool for every non-trivial design question, the operating-characteristics table had become the briefing-book centerpiece for every Type C meeting, and the ICH E9(R1) addendum had made the estimand specification a regulatory expectation rather than an implicit choice. The literature has lagged the practice: most published textbooks still organize their treatment around closed-form sample-size formulas and analytical operating characteristics, with the simulator as a supplementary appendix. This book inverts the organization. The simulator is the operative tool; the closed forms

are the sanity check.

The book has fourteen chapters and is best read in order. The experienced reader can dip into a specific chapter without losing the thread, but the chapters build on each other in ways that reward sequential reading.

Chapter 1 establishes the Monte Carlo foundation: why the parallel random-number generator's choice matters operationally, how to derive the Monte Carlo standard error and report it on every cell of every table, and how to reduce simulator variance by an order of magnitude through antithetic sampling, common random numbers, and control variates. Chapter 2 develops the operating-characteristics mindset that the rest of the book is in service of: the table is the regulatory artifact, the design is the vehicle that produces it, and a sponsor who has decided on the design before designing the table is doing trial design backward.

Chapters 3 and 4 take up the confirmatory-trial design questions for a Phase 3 survival endpoint. Chapter 3 contrasts the closed-form Schoenfeld prediction at proportional hazards with the realized operating characteristics under realistic non-proportional-hazards scenarios, where the unweighted log-rank loses eight to sixteen percentage points of power and the design must either accept the loss or change the analysis. Chapter 4 introduces the group-sequential framework with O'Brien-Fleming alpha-spending and the Cui-Hung-Wang weighting machinery for unblinded sample-size re-estimation, with explicit attention to the rpact-specific gotchas that have lost more than one trial team half a day's debugging.

Chapter 5 takes up estimands and intercurrent events under ICH E9(R1), making the difference between treatment-policy, hypothetical, composite, while-on-treatment, and principal-stratum estimands a quantitative statement rather than a philosophical one. The simulator reports the realized hazard ratio under each strategy across realistic switching scenarios; the choice between strategies is a clinical question (what does the regulator want to know about the drug?) supported by the simulator's numerical evidence.

Chapter 6 covers adaptive dose-finding in the Phase 1 setting, with operational comparisons of the four most widely-used algorithms (CRM, BOIN, mTPI-2, Keyboard) across three realistic dose-toxicity scenarios. The chapter's evidence is that the four algorithms produce essentially identical operating characteristics on realistic data and that the choice among them is operational rather than statistical. Chapter 7 develops multi-arm multi-stage designs with the multivariate-normal multiplicity correction; Chapter 8 develops Bayesian platform trials with EXNEX hierarchical borrowing and the family-wise error rate as the controlled quantity. Chapter 9 takes up external controls and the hybrid randomized-plus-external design, with the

propensity-weighting machinery, the commensurate-prior Bayesian dynamic-borrowing framework, and the e-value sensitivity analysis that regulators have come to expect.

Chapters 10 and 11 develop the operational and analytical machinery for survival endpoints. Chapter 10 handles event-driven analysis timing, the Anisimov-Fedorov accrual model, and the interim-update mechanism that narrows the predictive distribution on the final-analysis date by approximately fifty percent at the planned interim look. Chapter 11 develops the weighted log-rank class and the RMST framework as the regulatorily-acceptable alternatives to the unweighted log-rank under non-proportional hazards, with the Fleming-Harrington late-emphasis weight and the max-combo hedge as the modern defaults.

Chapter 12 takes up Bayesian decision rules at interim looks, with the predictive-probability-of-success machinery and its frequentist analog in conditional power. Chapter 13 develops the beta-spending framework for futility analysis, with binding and non-binding variants and the operational considerations of trial-team enforcement. Chapter 14 closes by assembling all of the preceding chapters' simulator outputs into the master operating-characteristics table, with the misspecification envelope as the sensitivity machinery and the reproducibility manifest as the framework's regulatory contract. Chapter 14 is the book's Rule 0 deliverable; the preceding thirteen chapters have each contributed evidence to it.

Four trials run through the book as worked examples; the four front-matter pages preceding this preface introduce them. Each chapter's simulator is calibrated against one of the four trials' parameters, and the chapter's operating-characteristics table is the trial's design deliverable for the chapter's specific regulatory question. The calibration block in the project metadata documents the parameter values and the published references against which they are calibrated; a reader adapting the book's framework to their own trial replaces the calibration block's values with their trial's specifics and re-runs the simulator without other modifications.

The book uses R for the borrowed-language simulator implementations, with the verbatim source listings reproduced in the appendix. The R code in the appendix is the research code; there is no separate research code that the published code abridges. A reviewer who installs R 4.3.3, runs renv::restore against the project's pinned package manifest, and runs any script from the appendix's listings on a clean working directory reproduces the figure whose sha256 is logged in the project metadata. The reproducibility is exact to within the floating-point arithmetic's machine-dependent specifics, and the audit trail from the briefing book's numerical claim to the simulator's specific

computation is the framework's regulatory contract.

The book is opinionated where the operational stakes are high. Where the literature offers three equally-defensible choices, the book picks one and defends it from regulatory experience. Where the literature offers one defensible choice and several common but indefensible choices, the book says so explicitly: naive sample-size re-estimation inflates type-I error to 0.07 in PROTON-3-like settings and is regulatorily unacceptable, the unweighted log-rank loses substantial power under realistic delayed-effect data-generating processes and should not be the default primary analysis for immune-checkpoint-inhibitor trials, the win-ratio test is inadmissible for pure survival endpoints and should be reserved for endpoint composites. The opinions are formed across thirty-one years of post-hoc protocol reviews and the consulting work that followed retirement. They are not always the consensus opinion of the methodological-statistics community, and where they diverge the book says where and why. The reader is welcome to disagree, with the caveat that disagreement is more useful when it is grounded in the simulator's evidence than when it is grounded in textbook intuition.

A final note on what the book is not. It is not a comprehensive treatment of Bayesian inference, of causal inference, of survival analysis, or of the methodology underneath any of the individual frameworks the chapters develop. The book's purpose is the operational integration of these frameworks into the trial-design discipline, and the chapters develop each framework only to the depth needed to read the operating-characteristics table that the framework produces. Pointers to the methodological literature are throughout the chapters; the reader who needs deeper treatment of any specific framework should consult the cited references. The book's value is in the practical application of the integrated framework to the regulatorily-relevant artifact; the deeper methodology is well-documented elsewhere and is the responsibility of the methodological literature rather than of this book.

I have written the manual that I wish had existed when I was a junior biostatistician staring at my first group-sequential boundary, and that I wish had existed thirty years later when I was negotiating a Type C meeting with FDA on an adaptive sample-size question no published reference covered. The book is the consolidated practical experience of a working career; it is not a research monograph, and it does not aspire to be one. The reader who finishes it should be competent at producing regulator-grade operating-characteristics tables for the trial-design questions of the late 2020s. The framework is the book's deliverable; the specific examples are illustrative.

Contents

Chapter 1

Monte Carlo Foundations for Trial Simulation

A clinical trial that is "powered at 0.90" is powered at 0.90 under a specific data-generating process, a specific test statistic, a specific decision rule, and a specific simulator. Change any one of those four, and the headline number changes by enough to matter. The job of this chapter is to make every word of that sentence operational — to define what Monte Carlo simulation means in the trial-design context, to choose a pseudo-random number generator suitable for parallel operating-characteristic runs, to fix a seed and stream discipline that survives a regulator's reproducibility audit, to derive the Monte Carlo standard error that must accompany every reported operating characteristic, and to work three variance-reduction techniques on a one-sample power problem with enough rigor that the wall-clock savings show up in the numbers.

None of this is novel methodology. The techniques are sixty years old; the implementations have been mature in R for a decade; the regulatory expectations have been written down in ICH E9 since 1998 and refined in ICH E9(R1) since 2019. What *is* novel, and what this chapter sets up for the remaining thirteen, is the discipline of treating the Monte Carlo simulator as the trial's primary regulatory artifact rather than as a back-of-envelope check on a sample-size calculation that was really done by formula. Every operating-characteristic table this book presents traces by `sha256` to a single named R script in the book's source repository; every reported probability carries a Monte Carlo standard error; every replicate count is justified by a written precision requirement; every random-number stream is reproducible to the digit by a regulator running the same code on a clean machine in 2031.

That discipline is the chapter's deliverable.

The four-trial portfolio this book uses for its worked examples — PROTON-3 (a Phase 3 oncology trial of an immune checkpoint inhibitor versus standard of care, with non-proportional hazards), VOLTA-1 (a Phase 1 dose-finding study in a first-in-human setting), AURORA (a Phase 2 basket trial with biomarker stratification across four substudies), and MERIDIAN (a rare-disease Phase 2 trial augmented with a propensity-weighted external control) — is introduced in detail in Chapter 2. This chapter uses two schematic examples instead: an exponential survival simulation for the random-number-generator diagnostic, and a one-sample mean test for the Monte Carlo standard error and variance-reduction work. We separate the machinery from the trial context deliberately. The reader who finishes the chapter should be able to write a small simulator, report its operating characteristics with calibrated uncertainty, and defend the reported numbers under scrutiny — before any of the regulatory complexity of the running portfolio enters the picture.

1.1 What "Monte Carlo" means in a trial-simulation context

The phrase *Monte Carlo* entered the statistical lexicon at Los Alamos in 1946 [1, 2], when Stanislaw Ulam, recovering from encephalitis and idly computing the probability that a Canfield solitaire hand would lay out, realized that the combinatorics he could not enumerate analytically he could approximate by repeated random play. Within two years von Neumann and Ulam had turned the trick into a workable algorithm for neutron transport calculations on the ENIAC; within three decades the same idea had spread to financial derivatives [3, 4], computer graphics, statistical mechanics, Bayesian inference [5], and — belatedly, and with less methodological coherence than it deserved — the design and analysis of clinical trials. The phrase is now ambiguous unless qualified, and the trial-simulation flavor of Monte Carlo has acquired its own characteristic obligations that this chapter is about.

1.1.1 What a trial-simulation Monte Carlo program contains

A Monte Carlo simulation of a clinical trial, in the regulatory sense this book uses the term, is a computer program with five components. Naming them

precisely matters because the regulatory artifacts the program produces are obligated against every one of the five.

Component one — the data-generating process. A probability model that produces a synthetic patient cohort, including baseline covariates, randomization arm allocation, intercurrent events (treatment discontinuation, rescue therapy, death from a competing cause), and the primary outcome of interest, whether continuous, binary, time-to-event, or composite. The data-generating process (DGP) is where the statistical content of the trial design lives. A PROTON-3 simulation, for example, generates 600 overall-survival times from a mixture of two Weibull populations matched to the published curves of a comparator immune-checkpoint trial, applies a non-proportional hazards effect under the proposed treatment, and overlays an accrual model calibrated to the participating sites' historical enrollment rates. Sometimes the DGP matches the null hypothesis (no treatment effect); sometimes it matches the protocol-stated alternative; sometimes it matches a perturbation the protocol authors have not formally hypothesized but a regulator might ask about (a slower accrual, a heavier-tailed survival distribution, a higher-than-expected dropout rate during the first six months). Producing operating characteristics across an envelope of DGPs is the work of Chapter 14; producing a single DGP that closes against published prior trials is the work of Chapter 2.

Component two — the trial protocol simulator. A piece of code that operates on the simulated cohort exactly as the real trial would: randomization scheme (stratified, permuted-block, dynamic allocation, response-adaptive), independent data monitoring committee triggers, group-sequential interim analyses, futility stops, dose-finding rules, sample-size re-estimation, primary analysis, sensitivity analyses. The protocol simulator is where the operational content of the trial lives. The same DGP can be paired with several protocol simulators to ask comparative questions — design A versus design B under matched DGP, group-sequential versus fixed-sample, binding versus non-binding futility — and the difference in their operating characteristics under common random numbers (§1.5) is what justifies the design choice.

Component three — the decision rule. A function that, given the trial output, returns one or more discrete outcomes per simulated trial. For confirmatory designs the decision rule is typically a single bit (reject the null at one-sided $\alpha = 0.025$, or fail to reject), but in adaptive designs and group-sequential designs it is richer: the look at which efficacy was declared, the look at which futility stopped the trial, the dose selected as the maximum tolerated dose, the arms surviving past interim 2, the patient subgroup car-

Chapter 3

Power and Sample Size by Simulation

The sample-size justification is one of the most read and least scrutinized pages of a Phase 3 protocol. The convention, settled in the early 1980s by Schoenfeld [32, 33] and Freedman [34] for survival endpoints and propagated through every regulatory biostatistics textbook since, is to state the required number of events under a target hazard ratio at a one-sided α and target power, in closed form, in a single equation. For PROTON-3 the convention produces the number 380: at a target hazard ratio of 0.72, one-sided $\alpha = 0.025$, and target power 0.90, the Schoenfeld formula requires 380 events at the primary analysis, and this number drives the trial's sample size of 600 patients, its accrual budget over 30 months, and its follow-up plan to a calendar date approximately 48 months after first patient in. The number is correct, in the sense that the closed-form derivation does what the derivation claims to do, and the number is also incomplete, in the sense that the derivation assumes proportional hazards and the realistic data-generating model for an immune-checkpoint-inhibitor trial does not have proportional hazards. The chapter is about that gap.

The gap matters because the trial's realized power — the probability that the log-rank test rejects the null at the analysis time — is a function of the actual hazard ratio trajectory under the trial's data-generating process, and the trajectory under the immune-checkpoint-inhibitor class of therapies is characteristically non-proportional. A delayed-effect pattern is well-documented across the class: the two arms' survival curves coincide for the first several months of follow-up while the treatment's immune-priming mechanism develops, then diverge after the delay produces a separated late-stage hazard ratio

that is typically smaller (more favourable) than the protocol's stated overall HR. A trial designed to 380 events under proportional hazards will, under a realistic delayed-effect data-generating process, deliver realized power not at the protocol-stated 0.90 but somewhere in the range 0.74 to 0.82, with the exact number depending on the length of the delay and the magnitude of the late HR. The protocol's headline 0.90 is an artifact of an assumption that does not hold, and the trial's actual operating characteristic — the number a reviewer would care about — is the simulated power under the realistic DGP, not the closed-form power under the assumed DGP.

This chapter constructs the simulation that produces the realistic power, in a form a reviewer can interrogate cell by cell. We start with the closed forms and their failure modes, work the PROTON-3 simulator under proportional hazards as a sanity check against Schoenfeld, then run the same simulator under three non-proportional hazards scenarios and report the operating-characteristics gap. The result is an OC table for PROTON-3 with four rows (PH and three NPH families) and three columns (log-rank, restricted mean survival, weighted log-rank), and a single recommendation that follows from the table. Chapter 4 will then ask what the design should do about the gap; Chapter 5 will ask what the estimand should be; this chapter only asks what the power is under the realistic DGP and shows the regulator the number.

3.1 The closed forms and when they fail

The Schoenfeld formula for the number of events required to test $H_0 : \mathrm{HR} = 1$ against $H_1 : \mathrm{HR} = \theta$ at one-sided significance α and target power $1-\beta$ under proportional hazards is

$$d \;=\; \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{p_1 p_2 \left(\log \theta\right)^2}, \tag{3.1}$$

where p_1 and p_2 are the per-arm randomization fractions ($p_1 = p_2 = 0.5$ for balanced randomization), θ is the target hazard ratio (e.g., 0.72 for PROTON-3), and $z_{1-\alpha}, z_{1-\beta}$ are the standard normal quantiles. The formula is exact under proportional hazards, exponential survival within each arm, balanced randomization, no censoring beyond administrative censoring at the analysis time, and the asymptotic distribution of the log-rank statistic. For PROTON-3 at $\theta = 0.72$, $\alpha = 0.025$ one-sided, $1 - \beta = 0.90$, balanced randomization, the formula gives

$$d \;=\; \frac{(1.96 + 1.282)^2}{0.5 \cdot 0.5 \cdot (\log 0.72)^2} \;=\; \frac{10.51}{0.25 \cdot 0.1077} \;=\; 390. \tag{3.2}$$

The protocol-stated 380 comes from rounding and a slight calibration to the expected censoring-rate adjustment Freedman [34] introduces; the two closed forms agree to within a couple of events at this design point. The exposition convention through the rest of this chapter is to refer to "the Schoenfeld estimate" as 380, the protocol-stated value, with the understanding that the precise derivation is the formula (3.1).

The Freedman variant of the closed form replaces the Schoenfeld asymptotic with a finite-sample correction that accounts for the actual censoring pattern; for the PROTON-3 censoring rate (approximately 37% at the analysis time, where the censored patients are those who have not yet had an event by the data cutoff), the Freedman estimate gives essentially the same number, 381. Lakatos [35] generalizes the closed form to handle stage-wise hazard ratios under non-proportional hazards, requires the analyst to integrate a stage-wise expression numerically, and is structurally a transitional artifact between closed-form and simulation; the regulatory practice today is to skip Lakatos and go directly to a Monte Carlo simulator for any non-PH question. Hasegawa [36] provides closed-form approximations under the weighted log-rank class for Fleming-Harrington weights, which the modal Phase 3 oncology trial uses under non-PH; the approximations are useful for back-of-envelope design but the regulator typically wants the simulated operating characteristics across the family of weights, not the closed-form approximation at a single weight.

The five places the closed forms fail, in increasing order of regulatory importance, are the following.

Non-proportional hazards. The Schoenfeld formula assumes a constant hazard ratio across follow-up. Under any non-PH scenario — delayed effect, crossing hazards, diminishing effect, accelerated failure under late censoring — the formula's d no longer corresponds to the events needed to achieve the target power. The direction of the error depends on the analysis: for a log-rank test applied to delayed-effect data, the formula overstates power (the realized power is lower than the closed-form claim); for a weighted log-rank with late-emphasizing weights applied to delayed-effect data, the formula understates power (the test's power is higher than the unweighted closed form). Either way the closed form does not match the simulator's output, and the trial's planning needs to be driven by the simulator.

Analysis-model mismatch. Even under proportional hazards, the closed form assumes the analysis is the score test of the log-rank statistic. A

Six months before the analysis. The FDA reviewer asks why the protocol's expected power of 0.90 doesn't match the simulator's 0.78 under realistic non-proportional hazards. The room goes quiet. The biostatistician opens a textbook and finds the Schoenfeld formula. The reviewer waits.

This is the book that ends that silence.

Across fourteen chapters built around four worked-example trials --- a Phase 3 immune-checkpoint comparison, a Phase 1 dose-finding study, a Phase 2 basket trial with hierarchical borrowing, and a rare-disease hybrid design with a propensity-weighted external control --- Dr. Ingrid Voss develops Monte Carlo simulation as the working biostatistician's primary tool for the regulatorily-relevant questions of the late 2020s. You will learn to size a trial against the data-generating process it actually produces, to choose a primary analysis the regulator will accept, to defend every cell of every operating-characteristics table at the Type C meeting, and to write simulators a third party can run on a clean machine five years later and reproduce your reported numbers to the digit.

The R code is in the appendix. The opinions are explicit. The discipline is what thirty-one years of regulatory negotiations have actually produced.