Improving the efficiency of A and D optimal designs for dose response models

Srichand Jasti

A Dissertation
Presented to the Faculty of
University of North Texas Health Science Center
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Biostatistics and Epidemiology
Adviser: Dr. Rajesh Ranjan Nandy

June, 2021

DocuSign Envelope ID: B17C45AE-5B7F-40B2-AA09-1B67C89D57E7

IMPROVING THE EFFICIENCY OF A AND D OPTIMAL DESIGNS FOR DOSE RESPONSE MODELS

SRICHAND JASTI

A DISSERTATION PRESENTED TO THE FACULTY OF UNIVERSITY OF NORTH TEXAS HEALTH SCIENCE CENTER IN CANDIDACY FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DocuSigned by:
Dr. Rajesh Nandy 6/23/2021
EFED35D5E94A44E...
Rajesh Nandy, PhD
Associate Professor and Committee Chair

DocuSigned by:
Md Ubydul Haque 6/23/2021
1C56F65E518A451...
Ubydul Haque, PhD
Associate Professor and Committee Member

DocuSigned by:
6/23/2021
A0C6D7DA087E482...
Subhash Aryal, PhD
Associate Professor and Committee Member

DocuSigned by:
Tracy E Barnett, PhD 6/23/2021
1F2B0EDEDA474CA...
Tracy Barnett, PhD
Interim Chair, Dept. of Biostatistics & Epidemiology

DocuSigned by:
Dennis Thombs 6/23/2021
EBAEA35F26BD400...
Dennis Thombs, PhD
Dean, UNTHSC School of Public Health

© Copyright by Srichand Jasti, 2021.

All rights reserved.

Abstract

This doctoral thesis investigates approaches to improve efficiency of A and D optimal designs for the two parameter logistic model.

Aim 1: For A-optimality, by virtue of Cramér–Rao bound, the trace of the inverse of Information matrix for the parameters serves as a lower bound for the sum of variances of the estimators and the bound is attained asymptotically. Hence, asymptotically, A-optimality is achieved by minimizing the trace of the inverse of the Information matrix. For non-linear models, Cramér–Rao bound is crude for finite samples and hence the asymptotic solution can be very different from the design that minimizes the sum of variances. We explore the validity of the asymptotic solution by directly minimizing the sum of variances using numerical methods in a restricted search space. We demonstrate that even in a very restrictive search space of point symmetric designs, the theoretical solution is half as efficient for a sample size of 100. Further improvement can be achieved by relaxing the restriction of the solution being point symmetric.

Aims 2 & 3: The solution to A and D optimal designs for the logistic model depend on the unknown parameters of the model. Therefore, to obtain an optimal design the experimenter must inform the design based on some prior knowledge, or a guess, of the unknown parameters. This is a severe limitation on the ability to identify an optimal design especially when there is little prior information to inform the guess. Here we explore the use of a two-stage A-optimal design for finite samples and three-stage D-optimal design for large samples to mitigate the loss in efficiency which may arise due to poor guess values. We demonstrate that while two-stage finite sample model results in gain in efficiency with small sample sizes at 70% allocation to the first stage. The three-stage D optimal design is shown to be almost always better than the single stage and the corresponding two-stage design.

Acknowledgements

I express my sincere gratitude to the many people who have served as ballast throughout a journey that I now consider the pinnacle of my personal accomplishment.

Foremost, Dr. Rajesh Nandy – thank you for the years of support and hand-holding. Without your help and patience, this would not be possible. Your thoughtful insights kept the research work in focus and kept me from straying too far. My committee members, past and present, Dr. Sumihiro Suzuki, Dr. Subhash Aryal, and Dr. Ubydul Haque – thank you for your timely reviews, advice, and motivation.

Dr. William Dudley, Dr. David Kramer, Dr. Hyung-Woo Kim, Dr. Brian Wiens, Dr. Ken Sullivan each of you influenced my career interests and fostered an environment in which I was able to explore, experiment, learn, and pursue a doctoral degree while also working at Alcon.

Dedicated to my family - Soumya, Kavya, Veda & Lekha.

Contents

Chapter 1

Introduction

Statistical methodology is widely employed in the design and analysis of experiments and is a corner stone of scientific investigation. Historically, methodological research has focused extensively on addressing the properties of experiments that aim to compare outcomes between groups. In such experiments, the conditions of the experiments are known ahead of time and the goal is to evaluate the effect of a co-variate, such as treatment, on a response variable. The choice of design is motivated by a need to control the type I and type II errors through the choice of an appropriate sample size. Whereas this continues to be a fertile and important area of research, another important type of experiments aimed to describe and quantify the relationship between a stimulus and resulting response has gained much attention in recent years. While the sample size for such studies is usually based on cost or other practical considerations, important unknowns pertaining to the experimental design - such as the number of levels of the co-variate to study, the particular levels to study, and the proportion of the total sample size to be allocated to each of the chosen levels must be determined before embarking on the conduct of the experiment. A metric, such as a statistical criterion, is needed to choose among the potentially infinite number of possible designs. Such a criterion can aid in evaluating potential designs and help identify a design that is best. It can also help to evaluate the loss in efficiency in situations where a compromise must be made, e.g. when a particular design is not feasible due to other practical limitations. The theory of optimal designs considers this problem formally and is concerned with identifying an 'optimal design' with respect to a chosen statistical criterion.

1.1 Development of optimal design theory

In early 20$^{\text{th}}$ century, the use of statistics in design and conduct of experiments was pioneered by Sir R. A. Fisher and exemplified by the famous lady tasting tea experiment detailed in his book, *The design of experiments*, Fisher (1949). His work established the notion of the *null*-hypothesis and the use of probability theory to disprove the *null*. Soon thereafter the foundational principles from Fisher's work were adopted for application in several fields. The applications to industrial engineering in particular lead to the development of Response Surface Methodology (RSM) aimed to best characterize relationships between input and output variables, Box & Wilson (1951). By mid-century, Kiefer's initial investigations into optimal design theory also began and the development of the General Equivalence Theorem had advanced the estimation methods for linear models (Kiefer (1959); Kiefer & Wolfowitz (1960); Kiefer (1961)). The extensions to non-linear models followed soon thereafter White (1973) and specific applications of the optimal design theory to the two parameter logistic model as they relate to dose-response studies in the pharmaceutical industry were also developed Heise & Myers (1996); Wu (1985).

For a more complete review of the history and development of optimal design theory the reader is directed to the following books - Shah & Sinha (1989); Silvey (1980); Pukelsheim (1993); Fedorov & Leonov (2013).

By early 21 th century, the advances in computer technology and availability of cheap computational power have furthered several more advances specifically as they apply to generalized linear models.

1.2 Relevance to Public Health

Public health investigations and interventions aim to maximize the common good for populations by evaluating strategies to improve health. Experimental design is critical to unambiguously identify factors that can improve health. For any interventions involving a dose-response relationship, optimal design of experiments can provide a way to maximize the information obtained.

Several recent studies in the public health sphere have employed optimal designs in wide ranging fields. Fredrickson et al. (2005) applied optimal design methods in survey methodology to determine ways to maximize the response rate to a survey questionnaire. Through the use of an optimized contingent financial incentive they were able to realize an increase in response rate of 20-25 %. Klick et al. (2012, 2014); Cope (2019) considered the optimal design methods to study influenza transmission. They investigated the design parameters of sample size and follow-up intensity for estimation and comparison of secondary attack proportion of influenza.Santos et al. (2020) investigated the use of optimal methods in antibiotic research to map the response surface of Eosin Y concentration and irradiation time on Staphylococcus aureus counts.

To illustrate the application of optimal design of experiments to public health problems in a contemporaneous example, consider a public health department aiming to increase COVID-19 vaccination rate through a monetary compensation program. One way to approach this problem is to conduct an experiment by providing several levels of compensation to groups of subjects and ascertaining the rate of vaccination at each compensation level. To conduct such an experiment the following questions must be adequately considered:

- How many compensations levels to study?
- What is the compensation at each level?
- Given an overall sample size, what proportion of subjects should be allocated to each level?

In practice, the answer to each of the above question is based on the best judgment of the investigator and available historical evidence from literature and other sources. This is subject to the experience of each investigator and does not provide a quantitative way to evaluate the impact of varying a factor on the design efficiency. The theory of optimal design of experiments provides answers each of these questions based on optimization of a chosen statistical criterion.

While the example described here is simple and avoids the discussion of myriad other factors which may affect the choice of the final design, it powerfully illustrates a practical application to public health problems. The techniques presented in this thesis can be expanded and applied to incorporate other factors and more parameters.

1.3 Optimality Criterion

Optimal designs aim to maximize the amount of information obtained from an experiment. To this end, and owing to the inverse relationship between variability and statistical information, it is helpful to think about the goal of the design in terms of minimizing variability of the parameter estimates. When more than one parameter is required to characterize a relationship, e.g. intercept and slope in a two-parameter regression model, a carefully chosen function of the information matrix may serve as the statistical criterion. The chosen criterion, based on the inferential goals of the study, may represent minimizing the variability of a single parameter, or a combination of parameters, or minimizing the generalized variance of the parameter estimates. Jack Kiefer is credited with developing the alphabetical notations to describe the optimality criteria Kiefer (1959). The focus of this work will be two specific criteria that are widely employed in the design of optimal experiments, A-optimality and D-optimality. A-optimality seeks to minimize the sum of the variances of parameters and D-optimality seeks to maximize the determinant of the information matrix and hence minimize the generalized variance. Prior theoretical work primarily addressed the optimal design problem for the linear models, but in recent years has been extended to non-linear models through recent developments in the theory of generalized linear models (GLMs).

1.4 Scope of work

This doctoral thesis investigates approaches to improve efficiency of optimal design of experiments for the two parameter logistic model.

While the theoretical A-optimal solution based on the Cramér–Rao bound is attained asymptotically, it is hypothesized that direct minimization of the A-optimality criterion using numerical methods can lead to identification of alternate designs with improved efficiency. It is further hypothesized that by expanding the search outside of the symmetric design space further efficiency can be gained.

It is also noted that the information matrix for linear models is independent of the unknown parameters of the relationship that the experiment is intending to model. This is not the case with non-linear models and, for such models, the experimenter informs the design based on some prior knowledge, or a guess, of the unknown parameters to obtain an optimal design. This is a severe limitation on the ability to identify an optimal design especially when there is little prior information to inform the guess. One approach to address this problem is through the use of multi-stage models. Nandy & Nandy (2015), investigated the properties of a two-stage design and demonstrated that when a proportion of the overall sample is utilized to inform and update the initial guess, the resulting four point design results in improved efficiency compared to the theoretical design and that this design performs better than using a single stage framework, on average. However, it is also recognized that when the initial guess for the first stage is very far from the true values the overall efficiency reduces considerably.

This leads to the main topics of this thesis:

- Specific Aim 1: To explore the validity of the A-optimal asymptotic solution by directly minimizing the sum of variances using numerical methods in a limited search space for finite sample sizes using a two-parameter logistic model.

- Specific Aim 2: To explore the properties and efficiency of a two-stage finite sample model employing the A-optimality criterion.

- Specific Aim 3: To explore the properties and efficiency of a three-stage model employing the D-optimality criterion.

A brief description of the following chapters of the thesis follows. Chapter 2 provides an introduction to statistical theory relating to GLMs, parameter estimation, and optimality criteria. Chapter 3, 4, and 5 present results from the investigation of specific aim 1, 2, and 3 respectively. Chapter 6 summarizes the findings from the investigations outlined in this thesis and describes future avenues of research.

Chapter 2

Theory

2.1 Design notation

We introduce some notation and a formal definition of the optimal design problem.

As stated previously, for a given underlying model, the choice of a design comprises of three main decisions namely, number of doses (or levels), magnitude of each dose, and the number of subjects at each dose.

Formally, an experimental design, $\mathcal{D}$, is defined as $\mathcal{D} = \{(x_i, \xi_i), i = 1, 2, \ldots, m\}$, where x_i and ξ_i are the i^{th} dose (or level) of stimulus and weight at that stimulus respectively such that $\sum_{i=1}^{m} \xi_i = 1$ and all $\xi_i > 0$, implying a continuous setting. We also assume that the dose range is $0 < x_i < \infty$. A design is represented as follows

$$\mathcal{D} = \begin{pmatrix} x_1 & x_2 & \ldots & x_m \\ \xi_1 & \xi_2 & \ldots & \xi_m \end{pmatrix}$$

The problem definition involves the identifying the number of experimental levels m, the dose levels x_i (also referred to as design points), and the proportion of sample size allocated at each dose level ξ_i (also referred to as weights), also known as weights at each design point, such that a chosen statistical criterion is optimized and amount of information obtained from the experiment is maximized.

2.2 Logistic regression

We begin by introducing the logistic model for binary data. Let $Y_1, Y_2, \ldots, Y_n$ be binary responses for n subjects and x_{i1}, x_{i2},..., x_{ip} be the p regression variables for subject i. The class of General Linear Models (GLMs) for this data can be formulated as

$$prob(Y_i = 1|x_i) = P(\boldsymbol{X_i}^T \boldsymbol{\theta})$$

Where, $\boldsymbol{X_i} = (1, x_{i1}, \ldots, x_{ip})^T$, $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)^T$ is the vector of unknown parameters, and $P(x)$ is a cumulative distribution function (*cdf*). For the logistic model the *cdf* is defined to be $P(x) = \frac{1}{1+e^{-x}}$. The likelihood function for $\boldsymbol{\theta}$ can be written as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} P(\boldsymbol{X_i}^T \boldsymbol{\theta})^{Y_i} (1 - P(\boldsymbol{X_i}^T \boldsymbol{\theta}))^{(1-Y_i)}$$

The likelihood equations are

$$\sum_{i=1}^{n} \boldsymbol{X_i} \frac{[Y_i - P(\boldsymbol{X_i}^T \boldsymbol{\theta})] P'(\boldsymbol{X_i}^T \boldsymbol{\theta})}{P(\boldsymbol{X_i}^T \boldsymbol{\theta})(1 - P(\boldsymbol{X_i}^T \boldsymbol{\theta}))} = 0$$

The maximum likelihood estimates of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ are obtained by solving the non-linear equations numerically. The Fisher information matrix can be obtained as follows

$$E\left(-\frac{\partial^2 lnL(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) = \sum_{i=1}^{n} \boldsymbol{I}_{\boldsymbol{X}_i} = \frac{P'(\boldsymbol{X_i}^T \boldsymbol{\theta})^2}{P(\boldsymbol{X_i}^T \boldsymbol{\theta})(1 - P(\boldsymbol{X_i}^T \boldsymbol{\theta}))}$$

$\boldsymbol{I}_{\boldsymbol{X}_i}$ is the information matrix for $\boldsymbol{\theta}$ at a the design point $\boldsymbol{X_i}$.The inverse of $P(x)$ is known as the link function for these GLMs and for logitic model, this link function corresponds to the *logit*. For a two-parameter logistic model, the GLM can simplified to

$$prob(Y_i = 1|x_i) = \frac{1}{1 + e^{-(\alpha+\beta x_i)}} = \pi(x_i)$$

Where α and β are the unknown parameters and $\beta > 0$. This can also be written as

$$\log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = logit[\pi(x_i)] = \alpha + \beta x_i$$

The variance of Y_i depends of the mean response μ_i. For the binomial response model $V(Y_i) = r_i\mu_i(1 - \mu_i)$, where r_i is the number of observations at x_i. It is worthwhile to note that the link function for a normally distributed response is identity and hence the variance $V(Y_i) = \sigma^2$ is a constant. Note that both A and D optimality criteria are functions of the information matrix which, in turn, is inversely related to the variance. Thus, it is easy to see that the solution for the logistic model optimal design depends on the true unknown parameters through μ_i, while the optimal design solution for the standard linear model is independent of the unknown parameters.

2.3 Information matrix

The information matrix for the joint estimation of α and β for a two-parameter logistic model is obtained as follows. The likelihood for the i^{th} observation

$$L_i(\alpha, \beta|x_i) = \begin{cases} \pi(x_i) & \text{if } y_i = 1 \\ 1 - \pi(x_i) & \text{if } y_i = 0 \end{cases}$$

Alternately, the above may be combined into

$$L_i(\alpha, \beta|x_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$$

The log-likelihood is

$$l_i(\alpha, \beta | x_i) = (1 - y_i) \log(1 - \pi(x_i)) + y_i \log(\pi(x_i))$$

Ignoring the constant term, the joint log-likelihood is

$$\begin{aligned} l(\alpha, \beta | x_1, \ldots, x_n) &\propto \sum_{i=1}^{n} (1 - y_i) \log(1 - \pi(x_i)) + (y_i) \log(\pi(x_i)) \\ &= \sum_{i=1}^{n} y_i \log \frac{\pi(x_i)}{1 - \pi(x_i)} + \log(1 - \pi(x_i)) \\ &= \sum_{i=1}^{n} [y_i(\alpha + \beta x_i) - \log(1 + e^{\alpha + \beta x_i})] \end{aligned}$$

The information matrix for the two-parameter logistic model is a 2 x 2 matrix, with the following elements

$$I(\alpha, \beta) = \begin{pmatrix} -E[\frac{\partial^2 l(\alpha,\beta)}{\partial \alpha^2}] & -E[\frac{\partial^2 l(\alpha,\beta)}{\partial \alpha \partial \beta}] \\ -E[\frac{\partial^2 l(\alpha,\beta)}{\partial \alpha \partial \beta}] & -E[\frac{\partial^2 l(\alpha,\beta)}{\partial \beta^2}] \end{pmatrix}$$

Substituting the second derivatives in the information matrix

$$I(\alpha, \beta) = \begin{pmatrix} \sum_{i=1}^{n} \frac{e^{-(\alpha+\beta x_i)}}{(1+e^{-(\alpha+\beta x_i)})^2} & \sum_{i=1}^{n} x_i \frac{e^{-(\alpha+\beta x_i)}}{(1+e^{-(\alpha+\beta x_i)})^2} \\ \sum_{i=1}^{n} x_i \frac{e^{-(\alpha+\beta x_i)}}{(1+e^{-(\alpha+\beta x_i)})^2} & \sum_{i=1}^{n} x_i^2 \frac{e^{-(\alpha+\beta x_i)}}{(1+e^{-(\alpha+\beta x_i)})^2} \end{pmatrix}$$

For a design with m levels, each with $\xi_1, \xi_2, \ldots, \xi_m$ proportion of the total sample, the information matrix may be written as

$$I(\alpha, \beta) = \begin{pmatrix} \sum_{i=1}^{m} \xi_i \frac{e^{-(\alpha+\beta x_i)}}{(1+e^{-(\alpha+\beta x_i)})^2} & \sum_{i=1}^{m} \xi_i x_i \frac{e^{-(\alpha+\beta x_i)}}{(1+e^{-(\alpha+\beta x_i)})^2} \\ \sum_{i=1}^{m} \xi_i x_i \frac{e^{-(\alpha+\beta x_i)}}{(1+e^{-(\alpha+\beta x_i)})^2} & \sum_{i=1}^{m} \xi_i x_i^2 \frac{e^{-(\alpha+\beta x_i)}}{(1+e^{-(\alpha+\beta x_i)})^2} \end{pmatrix}$$

2.4 D-Optimality criterion

The D-optimality criterion seeks to minimize the generalized variance of the parameter estimates or, alternatively, maximize the determinant of the information matrix, there by maximizing the differential Shannon information content of the parameter estimates. From the information matrix, the determinant is computed as follows

$$\beta^2|I(\alpha,\beta)| = \Big[\sum_{i=1}^{m}\xi_i\frac{e^{-a_i}}{(1+e^{-a_i})^2}\Big]\Big[\sum_{i=1}^{m}\xi_i a_i^2\frac{e^{-a_i}}{(1+e^{-a_i})^2}\Big] - \Big[\sum_{i=1}^{m}\xi_i a_i\frac{e^{-a_i}}{(1+e^{-a_i})^2}\Big]^2$$

Where $a_i = \alpha + \beta x_i$. It is well known that the theoretical D-optimal solution is a two-point solution that is point-symmetric and weight-symmetric. That is, the D-optimal solution has the following form

$$\mathcal{D}_D = \begin{pmatrix} x_1 & x_2 \\ 0.5 & 0.5 \end{pmatrix}$$

Where, x_1 and x_2 are the design points (levels) and the total sample size is allocated equally to each of these design points. The design points are obtained by solving $-c_D = \alpha + \beta x_1$ and $c_D = \alpha + \beta x_2$. The D-optimality criterion is maximized at the critical value $c_D = 1.5434$. In practice, α and β are unknown and guess values are often used. The derivation of the D-optimal solution is detailed in several references Minkin (1987); Khan & Yazdi (1988); Sitter & Wu (1993); Mathew & Sinha (2001).

2.5 A-Optimality criterion

The A-optimality criterion aims to minimize the sum of the variances of the estimated parameters. That is, for the two-parameter logistic model, the A-optimality criterion seeks to minimize $Var(\hat{\alpha}) + Var(\hat{\beta})$. This criterion can be estimated from the information matrix as the trace of the inverse of the information matrix.

$$Var(\hat{\alpha}) + Var(\hat{\beta}) \geq \frac{\sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}\left(1 + \frac{(a_i-\alpha)^2}{\beta^2}\right)}{|I(\alpha, \beta)|}$$

Where $a_i = \alpha + \beta x_i$, and equality is only attained for large sample sizes.

The solution to the A-optimal design was first postulated by Mathew & Sinha (2001) under restricted conditions and later Yang (2008) established more rigorously. The A-optimal design is known to point-symmetric but not weight symmetric.

$$\mathcal{D}_A = \begin{pmatrix} x_1 & x_2 \\ \xi_1 & \xi_2 \end{pmatrix}$$

Again, as with D-optimality, x_1 and x_2 are the design points and are obtained by solving $-c_A = \alpha + \beta x_1$ and $c_A = \alpha + \beta x_2$. For the A-optimal solution, c_A is the positive solution of following quadratic equation

$$\frac{c_A^2 - \alpha^2 - \beta^2}{\sqrt{\beta^2 + (c_A + \alpha)^2} + \sqrt{\beta^2 + (c_A - \alpha)^2}} = 1 + \frac{c_A(1 - e^{c_A})}{1 + e^{c_A}}$$

The proportion of sample size allocated to the first design point is obtained by

$$\xi_1 = \frac{\sqrt{\beta^2 + (c_A + \alpha)^2}}{\sqrt{\beta^2 + (c_A + \alpha)^2} + \sqrt{\beta^2 + (c_A - \alpha)^2}}$$

Also, $\xi_2 = 1 - \xi_1$. For a more complete treatment, see Yang (2008).

Chapter 3

Finite Samples in A-Optimal Designs

Sample size chosen for any given experiment is limited especially in health sciences, either due to ethical considerations or resource limitations. Therefore, a finite sample is a consideration that researchers often encounter in practice. While the D-optimal design is independent of sample size, the A-optimal design is not. As discussed in the previous chapter, an A-optimal design is one that minimizes the sum of variances of all estimated parameters. The expression from section 2.5 for the lower bound for the A-optimality criterion holds asymptotically by virtue of Cramér–Rao bound. For non-linear models, the Cramér–Rao bound is known to be crude with small samples and hence the asymptotic solution can be very different from the design that minimizes the sum of variances.

This insight offers an opportunity to improve upon the A-optimal design by directly minimizing the sum of variances of the estimated parameters in a given search space through the use of numerical methods. The resulting improvement can be quantified as a measure relative to the A-optimal criterion of the theoretical solution.

3.1 Methodology

Without loss of generality the true, unknown values for (α, β) are chosen to be $(\alpha_t = 1, \beta_t = 1)$ for all investigations in this chapter. The methods outlined here can be easily applied to other values by re-scaling the results from this investigation. The outline of the chapter is as follows:

1. We start by estimating the theoretical A-optimal design using the method outlined in section 2.5

2. Under a restricted search space of two-point, point-symmetric but not weight-symmetric designs, we searched for designs which improve upon on the theoretical estimate for finite samples under the following conditions:

   - First, by fixing the doses to those determined by the theoretical solution, and searching for a proportion that minimizes the A-optimality criterion

   - Next, by fixing the proportion to the theoretical solution, and searching for point-symmetric doses that minimizes the A-optimality criterion

   - Finally, by conducting an exhaustive grid search in the aforementioned restricted space.

3. We complete the investigation by relaxing the point-symmetric restriction, and conducting an exhaustive grid search.

3.2 Performance of the theoretical A-optimal design

For $(\alpha_t = 1, \beta_t = 1)$, the theoretical A-optimal solution is identified to be $c_A = 1.482$ and $\xi_1 = 0.29$. The design points x_1 and x_2 are obtained by solving $-c_A = \alpha_t + \beta_t x_1$

and $c_A = \alpha_t + \beta_t x_2$. Therefore, the A-optimal design is

$$\mathcal{D}_A = \begin{pmatrix} -2.482 & 0.482 \\ 0.29 & 0.71 \end{pmatrix}$$

For this design, the magnitude of the A-optimality criterion employing the asymptotic solution is estimated as the trace of the inverse of the information matrix, standardized by the sample size of the experiment. That is,

$$A_{opt} = \frac{tr(I(\alpha_t, \beta_t)^{-1})}{n}$$

These estimates are considered to be reference values and indicate the best performance of the optimal design if asymptotic results hold. Next, to investigate the finite sample properties of this design under real experimental conditions, the following algorithm was employed and the A-optimal criterion was estimated directly.

1. Given a design, the probability of observing a successful outcome at each dose can be estimated using

$$p_1 = \frac{1}{1 + \exp\left(-1 + \frac{c_A - \alpha}{\beta}\right)} \text{ and } p_2 = \frac{1}{1 + \exp\left(-1 + \frac{-c_A - \alpha}{\beta}\right)}$$

2. For a given sample size, say n, random samples of size $n_1 = n * \xi_1$ and $n_2 = n * (1 - \xi_2)$ are generated representing the number of samples at each of the design points, i.e. low dose x_1 and high dose x_2 respectively, from a binomial distribution.

3. Next, a logistic regression model is fit to the resulting data set, with the random sample as the outcome variable and the design points as the input variable, to estimate the unknown parameters $\hat{\alpha}$ and $\hat{\beta}$

Sample Size (n)	c_A	x_{1D}	x_{2D}	ξ_1	A_{opt}	A^*_{opt}
20	1.482	-2.482	0.482	0.29	0.54	29.15
40	1.482	-2.482	0.482	0.29	0.27	4.84
60	1.482	-2.482	0.482	0.29	0.18	1.31
80	1.482	-2.482	0.482	0.29	0.14	0.57
100	1.482	-2.482	0.482	0.29	0.11	0.25
300	1.482	-2.482	0.482	0.29	0.04	0.04
1000	1.482	-2.482	0.482	0.29	0.01	0.01

Table 3.1: Theoretical and direct estimates of A-Optimality criterion for asymptotic solution at $(\alpha_t = 1, \beta_t = 1)$

4. Steps 2 and 3 are repeated for a large number of iterations $(25,000)$, and estimated parameters are saved for each iteration.

5. The direct estimate of A-optimality at the given design is obtained by calculating $A^*_{opt} = Var(\hat{\alpha}) + Var(\hat{\beta})$

6. Step 5 is repeated for each sample size and the results are presented in Table 3.1

It is evident from Table 3.1 that $A_{opt} \approx A^*_{opt}$, only for large sample sizes.

3.3 Efficiency of A-optimal designs

Section 3.2 outlines the procedure for directly estimating the A-optimality criterion through simulations. In order to investigate the improvement that can be achieved by directly minimizing the criterion under various assumptions, a search space was set-up for $c_{ASearch}$ ranging from 0.1 to 2.0 in 0.05 increments. For $\xi_{1Search}$, the range was set-up to be 0.1 to 0.9 in 0.04 increments. For various sample sizes, simulations

Sample Size (n)	c_A	$\xi^*_{1Search}$	A^*_{Search}	$E(\%)$
20	1.482	0.13	25.00	14.25%
40	1.482	0.41	3.85	20.52%
60	1.482	0.41	0.62	52.78%
80	1.482	0.49	0.26	55.32%
100	1.482	0.41	0.14	43.72%
300	1.482	0.29	0.04	0.00%
1000	1.482	0.29	0.01	0.91%

Table 3.2: Optimal proportion ξ^*_{search} for low dose at various sample sizes for ($\alpha_t = 1, \beta_t = 1$)

were performed for every combination of $c_{ASearch}$ and $\xi_{1Search}$ and an estimate of A-optimal criterion (A^*_{opt}) was obtained. An improvement (or loss) in efficiency for each design was estimated as

$$E = \frac{A^*_{Search} - A^*_{opt}}{A^*_{opt}} * 100\%$$

where, A^*_{Search} is the direct estimate at each design in the search space.

First, to investigate the impact of choice of allocation to each design point alone, we fixed c_A to the theoretical value of 1.482 and performed a search for the proportion allocated to low dose $\xi^*_{1Search}$ which resulted in a design that minimized the A-optimality criterion directly. The results from these simulations are presented in Table 3.2.

Next, we investigate the impact of the design points alone by fixing the allocation ratio $\xi_1 = 0.29$ and searching for the optimal design points critical value $c^*_{ASearch}$. The results of this investigation are presented in Table 3.3. From Tables 3.2 and 3.3, we note that while the designs converge at large sample sizes, there is substantial efficiency to be gained at small to moderate sample sizes.

Sample Size (n)	$c^*_{ASearch}$	ξ_1	A^*_{Search}	$E(\%)$
20	1	0.29	17.88	38.68%
40	0.95	0.29	1.95	59.74%
60	0.65	0.29	0.49	62.53%
80	0.95	0.29	0.21	63.35%
100	1.25	0.29	0.14	42.11%
300	1.5	0.29	0.04	-0.89%
1000	1.55	0.29	0.01	0.00%

Table 3.3: Optimal $c^*_{ASearch}$ at various sample sizes for $(\alpha_t = 1, \beta_t = 1)$

Sample Size (n)	$c^*_{ASearch}$	$\xi^*_{1Search}$	A^*_{Search}	$E(\%)$
20	0.5	0.55	12.938	55.62%
40	0.7	0.55	0.681	85.94%
60	1.15	0.51	0.289	77.99%
80	1.15	0.59	0.171	70.16%
100	1.3	0.63	0.128	48.18%
300	1.45	0.67	0.037	0.00%
1000	1.45	0.71	0.011	0.00%

Table 3.4: Optimal $c^*_{ASearch}$ and $\xi^*_{1Search}$ at various sample sizes for $(\alpha_t = 1, \beta_t = 1)$

Next, in the class of point-symmetric two-point designs, we relax both components of the theoretical design. The results are presented in Table 3.4. For small to moderate sample sizes, the efficiency gained is more than that achieved by varying the design points or the allocation alone. Further, we note that at large sample size of 1000, the design converges to the theoretical design.

Sample Size (n)	$x^*_{1Search}$	$x^*_{2Search}$	$\xi^*_{1Search}$	A^*_{Search}	$E(\%)$
20	-0.1	-3	0.87	10.27	64.78%
40	-0.1	-1.5	0.59	0.60	87.59%
60	0.3	-1.7	0.59	0.25	81.19%
80	0.6	-1.8	0.67	0.16	72.25%
100	0.5	-1.9	0.67	0.12	50.20%
300	0.5	-2.3	0.67	0.04	0.00%
1000	0.6	-2.3	0.71	0.01	0.00%

Table 3.5: Optimal designs at various sample sizes for $(\alpha_t = 1, \beta_t = 1)$

Finally, we relax the restriction of point symmetric designs and conduct a full grid search to investigate if there is more efficiency to be gained. We vary each design point $x^*_{1Search}$ and $x^*_{2Search}$ independently, and also vary the allocation to low dose $\xi^*_{1Search}$, at various sample sizes. These simulations are highly resource intensive and efficiency gained must weighed against the cost and time to search the full grid. The results are summarized in Table 3.5. We note that while there is approximately 10% improvement at small sample sizes, at moderate sample sizes the improvement reduces to $2-5\%$. At large sample sizes, there is no efficiency to be gained.

Figures 3.1 and 3.2 present a level plot of the directly estimated A-optimality criterion for symmetric designs for various designs.

The critical values C_A of the symmetric weights is presented along the x-axis and the proportion allocated at low dose ξ_1 is presented along the y-axis for various sample sizes of interest $(40, 60, 80, 100)$. Lower values of the A-optimality criterion are better and represented by the pink region. We note that, as the sample size increases, the area of the pink region increases, i.e. a larger proportion of the designs in the search space result in closer values of the A-optimality criterion. Further, the directly estimated A-optimal design is farther away from the theoretical A-optimal

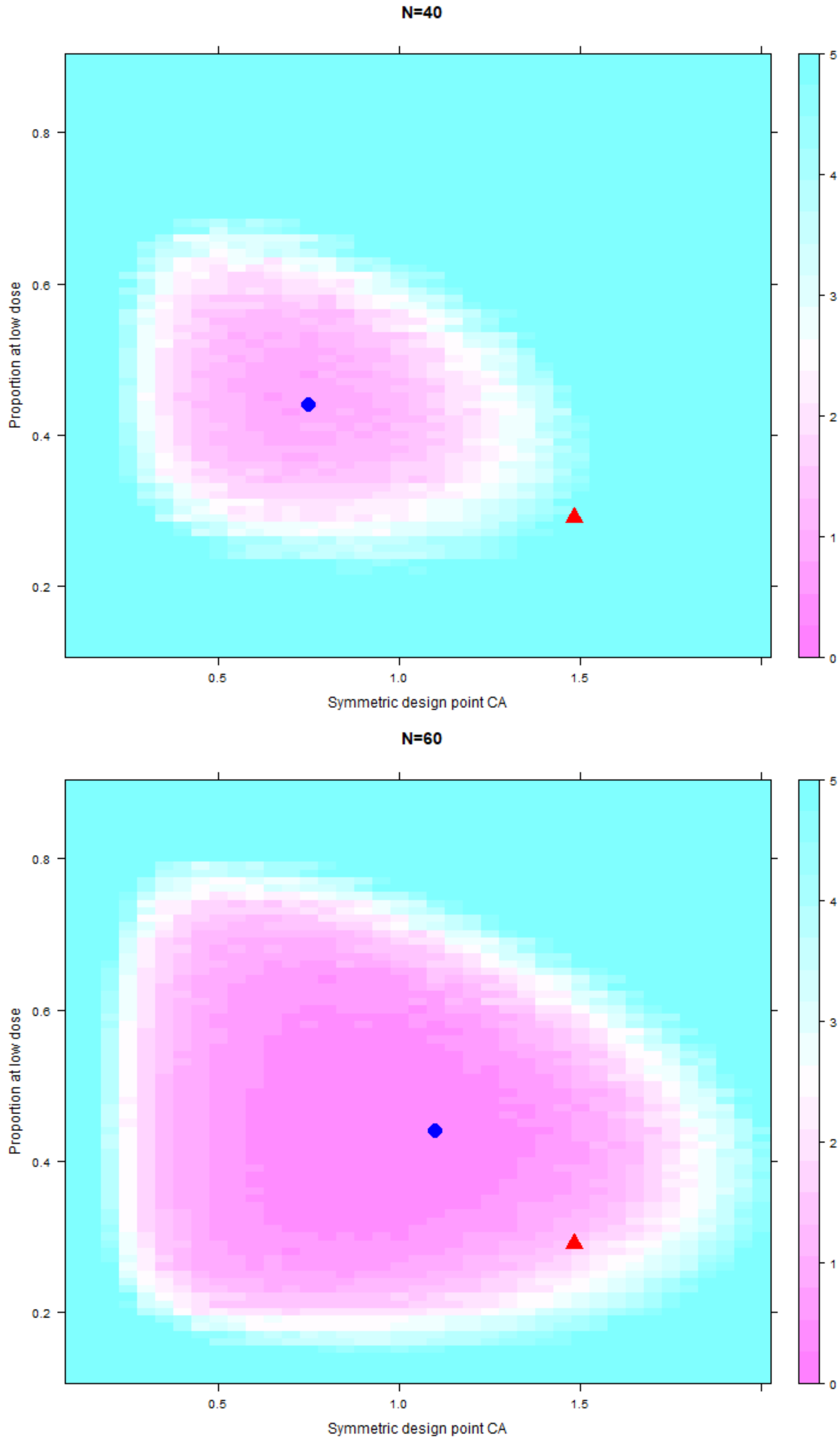

Figure 3.1: (●) Actual vs. (▲) Theoretical A-optimal design for ($\alpha_t = 1, \beta_t = 1$)

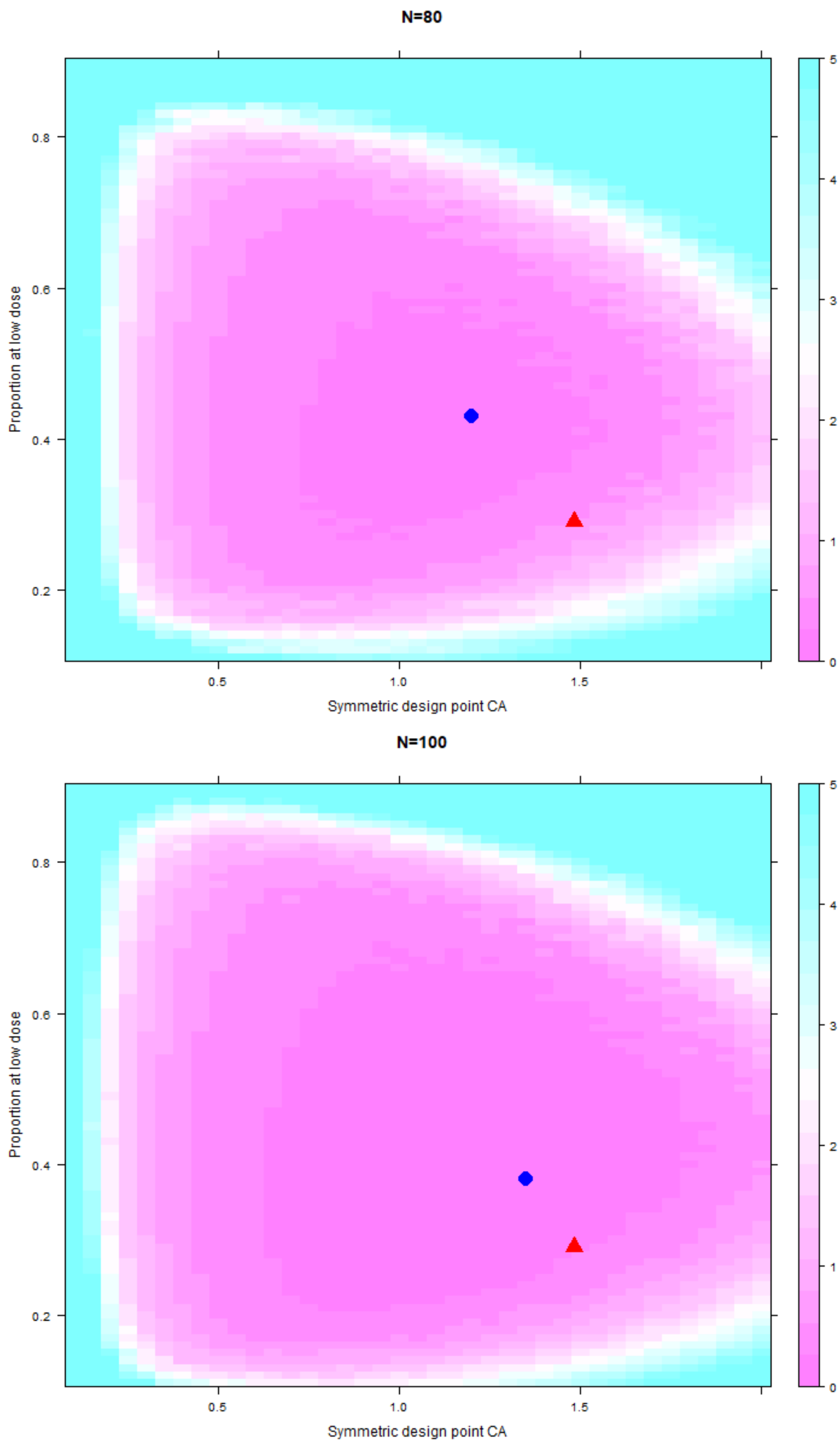

Figure 3.2: (●) Actual vs. (▲) Theoretical A-optimal design for $(\alpha_t = 1, \beta_t = 1)$

design at smaller sample sizes and approaches the theoretical design as the sample size increases.

3.4 Conclusions

The main conclusions from this investigation are presented below:

- Theoretical bounds are indeed achieved but for very large sample size
- Under finite samples the theoretical design does not minimize the sum of variances of the parameter estimates
- For finite samples significant improvement can be achieved by solving the optimality problem numerically even within the restricted space of symmetric designs
- Further improvement can be achieved by relaxing the point symmetric restriction – modest improvement observed at small sample sizes.

Chapter 4

Two-stage Finite Sample in A-Optimal Design

The preceding chapter established the importance and gain in efficiency that can be realized by employing direct minimization of sum of variances for a finite sample A-optimal design. This gain in efficiency is still constrained by the initial guess values of the unknown parameters, with poor initial guess leading to less improvement in efficiency. Nandy & Nandy (2015) demonstrated that, for large samples, a two-stage design leads to greater efficiency for both A and D-optimal designs. In this chapter, we investigate the applicability of this result to finite sample designs. We will evaluate whether the impact of poor initial guess values can be mitigated by employing a two-stage design in a finite sample setting by iterating the estimates of the unknown parameters. We will consider the impact of the total size of the experiment and the impact of the proportion of sample allocated to first stage.

4.1 Methodology

1. Consider an experiment of size N and an initial unknown parameter guess values of α_1 and β_1.

2. As with prior chapters, without loss of generality, the true values of the unknown parameters are assumed to be $(\alpha_t = 1, \beta_t = 1)$.

3. For a given experiment, the sample size is divided into two proportions such that $\sum \pi_i = 1$ where, $i \in \{1, 2\}$ and represents the stage of the experiment.

4. For the first stage, guess values of the parameters, (α_1, β_1) are employed to search for the finite sample A-optimal design as outlined in section 3.2.

5. The resulting design is employed to conduct an experiment (simulation) with sample size $N * \pi_1$.

6. Estimates of unknown parameters are obtained from the resulting data, say α_2 and β_2.

7. α_2 and β_2 are then employed as the best estimates for unknown parameter values for the next stage finite sample A-optimal design search as outlined in section 3.2.

8. The resulting design is employed to conduct an experiment (simulation) with the remaining sample of size $N * \pi_2$.

9. The simulated sample from stage I and stage II are combined and final estimates of the unknown parameters $\hat{\alpha}$ and $\hat{\beta}$ are obtained by performing a logistic regression.

10. Steps 5 through 9 are repeated a large number of times $(15,000)$ to directly estimate the A-optimal criterion $var(\hat{\alpha}) + var(\hat{\beta})$ is obtained.

4.2 Approach and Efficiency

We investigate this approach using two sample sizes, $N = 100$ and 200, with the proportion allocated to the first stage being varied from 0.3 to 0.7. The choice of sample sizes is based on findings from chapter 3 where sample sizes less than 300 are subject to the finite sample efficiency, while also allowing for adequate number of samples to be available at stage I to avoid singularity issues.

The efficiency of the two-stage finite sample A-optimal design is evaluated against the single stage finite sample A-optimal design which, in turn, has been demonstrated to be better than theoretical design.

$$E = \frac{A^*_{I-Stage} - A^*_{II-Stage}}{A^*_{I-Stage}} * 100\%$$

Due to the large number of simulations involved, the direct estimation approach is expensive in terms of computing resources and time. In a two-stage finite sample design, there are two direct estimations involved for each experiment, one at the initial search stage with the assumed guess values and another within the experimental simulations with the estimates from stage I to obtain the design to be implemented for the second stage. The number of simulations required for a complete characterization of the two-stage finite sample approach for a single combination of the chosen finite sample size and the proportion allocated to the first dose in the chosen search space is in the order of several hundred billion simulations. To circumvent a portion of this computation, a modified scaling approach is employed. In this approach, a look up table of designs and associated dose values are first generated by setting the initial guess values $\beta_1 = 1$ and varying $\alpha_1 \in (-50, 50)$. This lookup table can then be leveraged to scale the unknown parameter estimates from stage I to the form $(\frac{\alpha_1}{\beta_1}, \frac{\beta_1}{\beta_1}) = (\frac{\alpha_1}{\beta_1}, 1)$ and identify the corresponding dose values. Finally, these dose values are scaled by β_1 i.e. $(\frac{x_{ld}}{\beta_1}, \frac{x_{ud}}{\beta_1})$ and employed for the stage II experiment.

Even so, the number of simulations are still very large for each initial guess value. To evaluate this design, the efficiency gained is evaluated under two conditions

1. When the initial guess values (α_1, β_1) are far away from the true values $(\alpha_t = 1, \beta_t = 1)$ employed for the simulations

2. When the initial guess values (α_1, β_1) is identical to the true values $(\alpha_t = 1, \beta_t = 1)$ employed for the simulations

4.3 Results

Tables 4.1 and 4.2 present the results of the two-stage finite sample A-optimal designs for sample sizes of $N = 100$ and $N = 200$, respectively. Stage I A-optimality criterion is evaluated by searching for the finite sample A-optimal design under the assumed initial guess conditions and then conducting the simulated experiments with the full sample size allocated to the identified design. The two-stage design employs a portion of the full sample size to conduct the initial search and rest is employed to stage II of the design obtained from a second search using the estimates of the unknown parameters from the first stage.

In Table 4.1, we see that when the guess value is far away from the true parameter values, the gain in efficiency is $60 - 75\%$ based on the proportion allocated to the first stage. Higher proportion of the total sample size allocated to stage I of the design results in a higher efficiency or a lower loss in efficiency, when the when the initial guess values are close to the unknown true values.

Table 4.2 presents the results comparing the efficiency of the two-stage design to a single stage finite sample A-optimal design. We note that with a larger total sample size, the two-stage design performs poorly. Again, we note that the larger the proportion of the sample allocated to the first stage, the smaller the loss in efficiency.

N=100		Single stage design	Two stage design		E (%)
True values for experiment	Initial guess values	A-optimality criterion	Stage I proportion	A-optimality criterion	
(1.0,1.0)	(0.5, 0.5)	0.879	30%	0.886	- 0.75%
			40%	0.368	58.13%
			50%	0.221	74.82%
			60%	0.258	70.69%
			70%	0.234	73.33%
(1.0,1.0)	(1.0,1.0)	0.132	30%	0.886	-419.62%
			40%	0.293	-121.70%
			50%	0.202	-52.87%
			60%	0.162	-22.47%
			70%	0.150	-13.52%

Table 4.1: Efficiency of two-stage finite sample A-optimal design, $N = 100$

4.4 Conclusion

The two-stage finite sample A-optimal design is shown to be an improvement over the single stage design under specific conditions. With sample sizes of approximately $N = 100$ and miss-specified initial guess values, an increase in efficiency of $60-75\%$ can be obtained by employing the two-stage finite sample model. This gain in efficiency is not retained with larger sample sizes and the two-stage finite sample performs similar to the single stage finite sample A-optimal design.

N=200		Single stage design	Two stage design		E (%)
True values for experiment	Initial guess values	A-optimality criterion	Stage I proportion	A-optimality criterion	
(1.0, 1.0)	(0.5, 0.5)	0.069	30%	0.098	-42.65%
			40%	0.087	-26.64%
			50%	0.074	-7.71%
			60%	0.074	-7.71%
			70%	0.073	-6.26%
(1.0, 1.0)	(1.0, 1.0)	0.058	30%	0.067	-15.52%
			40%	0.062	-6.90%
			50%	0.060	-3.45%
			60%	0.060	-3.45%
			70%	0.059	-1.72%

Table 4.2: Efficiency of two-stage finite sample A-optimal design, $N = 200$

Chapter 5

Three-stage designs employing D Optimal Designs

Chapters 3 and 4 are concerned with properties of finite sample models when dealing with small sample sizes, in this chapter we explore potential increase in efficiency that can be gained in larger experiments through an iterative process. As outlined in Chapter 1, optimal designs for non-linear models are dependent on the unknown parameters of the model. An informed guess is frequently employed to circumvent this problem but there are inherent limitations to this approach.

1. In case of new products, little to no prior information may be available.

2. Due to subjectivity in investigator approach, there may be variability in choice of initial guess value with no way to evaluate the correct choice until after the experiment.

Nandy & Nandy (2015) demonstrated the improvement in efficiency when employing a two-stage design in lieu of a single stage design based on a guess value. Their work established that a two-stage design in which a portion of the sample is allocated to an initial design informed by guess values and the rest is allocated to the another design informed by the parameter estimates obtained from the initial stage performs at least as well as a single stage design when the guess value is close to the true value and significantly better, on average, when the guess deviates from the true value.

In this chapter, we explore whether additional efficiency can be obtained by employing a three stage D-optimal design and under which conditions it may be better than a two-stage design. We will consider the impact of the total sample size of the experiment and the impact of proportion of sample allocated to each stage.

5.1 Methodology

1. Consider an experiment of size N.

2. As with prior chapters, and without loss of generality, the true values of the unknown parameters are assumed to be $(\alpha_t = 1, \beta_t = 1)$.

3. For a given experiment, the sample size is divided into three proportions such that $\sum \pi_i = 1$ where, $i \in \{1, 2, 3\}$ represents the stage of the experiment.

4. For the first stage, guess values of the parameters, say (α_1, β_1), are employed to estimate the initial theoretical D-optimal design as outlined in section 2.4.

5. Given a design, the probability of observing a successful outcome at each dose can be estimated using

$$p_1 = \frac{1}{1 + \exp\left(-1 + \frac{c_D - \alpha_1}{\beta_1}\right)} \text{ and } p_2 = \frac{1}{1 + \exp\left(-1 + \frac{-c_D - \alpha_1}{\beta_1}\right)}$$

6. For the sample size chosen for the first stage, $n_1 = N * \pi_1$, random samples of size $n_1/2$ are generated representing the number of samples at each of the two design points, i.e. low dose x_1 and high dose x_2 respectively, from a binomial distribution.

7. Next, a logistic regression model is fit to the resulting data set, with the random sample as the outcome variable and the design points as the input variable, to estimate the unknown parameters $\hat{\alpha_1}$ and $\hat{\beta_1}$ at stage I.

8. The above estimates are employed as the guess parameter values for the next stage such that $\hat{\alpha_1} = \alpha_2$ and $\hat{\beta_1} = \beta_2$. With (α_2, β_2), the theoretical D-optimal design for the next stage is estimated.

9. Steps 5 to 7 are repeated with sample size stage II sample size of n_2 to obtain the stage II estimates $\hat{\alpha_2} = \alpha_3$ and $\hat{\beta_2} = \beta_3$. Note: The logistic regression model is fit to all samples from the first and the second stage simulations such that the total sample size is $n_1 + n_2$.

10. Steps 4 to 7 are repeated for a large number of simulations $(10,000)$ for each pair of initial guess values (α_1, β_1) and the average D-optimality criterion is calculated.

The above algorithm results in a three stage D-optimal design with six design points as below

$$\mathcal{D}_D = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ 0.5 * \pi_1 & 0.5 * \pi_1 & 0.5 * \pi_2 & 0.5 * \pi_2 & 0.5 * \pi_3 & 0.5 * \pi_3 \end{pmatrix}$$

The resulting three stage D-optimality criterion can be evaluated as a weighted average of individual two-point designs, as outlined in section 2.4 from each stage.

$$D_{III-Stage} = |\pi_1 * I(\alpha_1, \beta_1)_I + \pi_2 * I(\alpha_2, \beta_2)_{II} + \pi_3 * I(\alpha_3, \beta_3)_{III}|$$

5.2 Approach and Efficiency

We begin the investigation by defining three scenarios of sample size allocation to each stage. The first scenario starts with a small proportion of the overall sample size allocated to first stage with later stages being allocated with larger proportions. The second scenario allocates equal proportion to all stages and the final scenario employs the largest proportion to the first stage and and progressively smaller proportions for later stages. These scenarios are summarized in 5.1.

The search grid for initial guess parameters (α_1, β_1) is setup such that the true value $(\alpha_t = 1)$ and $(\beta_t = 1)$ are contained. We vary α_1 from -0.5 to 2.5 in steps of 0.1 and β_1 from 0.5 to 2 in steps of 0.1.

	Stage		
	I	II	III
Scenario 1	1/6	2/6	3/6
Scenario 2	2/6	2/6	2/6
Scenario 3	3/6	2/6	1/6

Table 5.1: Scenarios for proportion of sample allocated by stage

Finally, we repeat the investigation for several sample sizes $N = 150, 300, 600, 900$ and 1200. The sample sizes chosen ensure that each stage of the study has adequate sample size and does not run in to finite estimation issues.

The efficiency of resulting six point design can be compared the reference values of one-stage, two-point design outlined in 2.4 and the two-stage, four-point design identified in Nandy & Nandy (2015) using the the following expressions

$$E_{III-I} = \frac{D_{III-Stage} - D_{I-Stage}}{D_{I-Stage}} * 100\%$$

$$E_{III-II} = \frac{D_{III-Stage} - D_{II-Stage}}{D_{II-Stage}} * 100\%$$

Where,$D_{I-Stage}$ and $D_{II-Stage}$ are the reference D-optimality criteria for the one-stage and two-stage designs, respectively. The proportion allocation to first stage of the two-stage design was set to be the same as Stage I proportion in Table 5.1, with the rest of the sample allocated to the second stage. For the one-stage design the entire sample was allocated to the initial theoretical design based on guess values.

5.3 Results

Figures 5.1 to 5.5 present the gain in efficiency from a one-stage design to a three-stage design in the form of a heat map for various sample sizes under each of the allocation scenarios presented in 5.1. The initial guess values of the unknown parameters α and β are depicted along the x and y axis respectively. Each (x, y) coordinate on the grid represents the gain in efficiency obtained when employing a three stage design compared to a reference one-stage design. For the sake of legibility of the plots, the peak gain in efficiency is set to 100% with any higher values also set to 100%. Similarly, any values less than 0% i.e. cases in which three-stage design is worse than the one-stage design is set to 0%. Table 5.3 presents the actual minimum, media, and maximum values of efficiency gained at various sample sizes under each scenario without aforementioned limits.

Inspecting the figures and the tables, we observe that the three-stage design results in a gain in efficiency over a one-stage design in most of the search space except for when the guess values are very close to the true values. When the guess values are far from the true values, the three stage design can result in a substantial increase in the efficiency.

		Gain in Efficiency From One-Stage to Three-Stage		
		Min	Median	Max
N=150	Scenario 1	-20.6%	8.8%	1301.3%
	Scenario 2	-7.5%	18.5%	992.0%
	Scenario 3	-4.0%	17.9%	611.1%
N=300	Scenario 1	-7.5%	29.0%	1461.8%
	Scenario 2	-3.5%	29.8%	1080.2%
	Scenario 3	-2.0%	23.7%	664.2%
N=600	Scenario 1	-3.4%	39.3%	1566.5%
	Scenario 2	-1.7%	34.1%	1155.6%
	Scenario 3	-1.0%	26.4%	726.7%
N=900	Scenario 1	-2.3%	42.4%	1616.8%
	Scenario 2	-1.2%	35.7%	1204.7%
	Scenario 3	-0.7%	27.2%	783.9%
N=1200	Scenario 1	-1.7%	43.8%	1652.1%
	Scenario 2	-0.9%	36.3%	1246.6%
	Scenario 3	-0.5%	27.6%	830.2%

Table 5.2: Gain in efficiency from one-stage to three-stage design

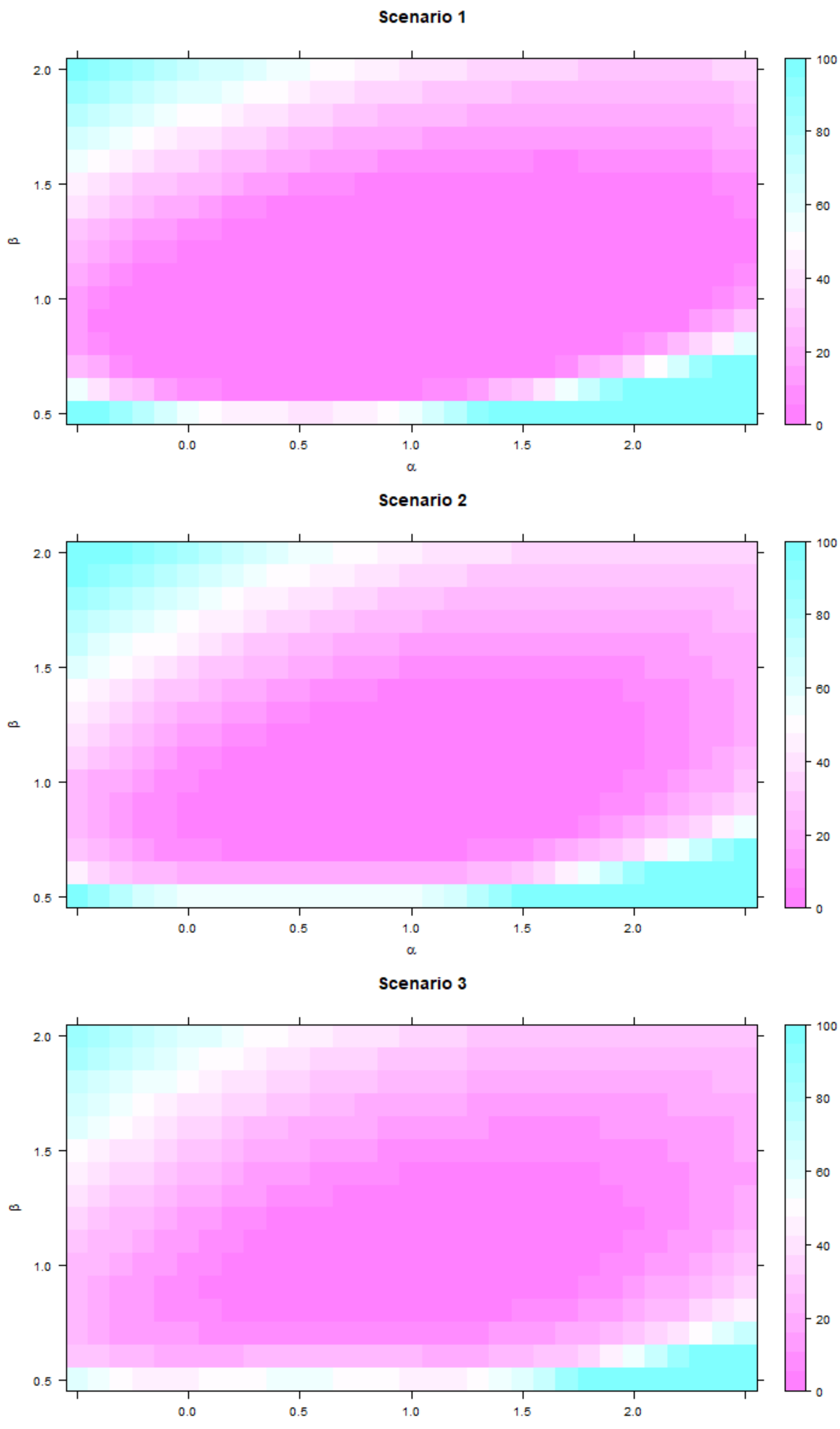

Figure 5.1: Gain in efficiency from one-stage to three-stage design, $N = 150$

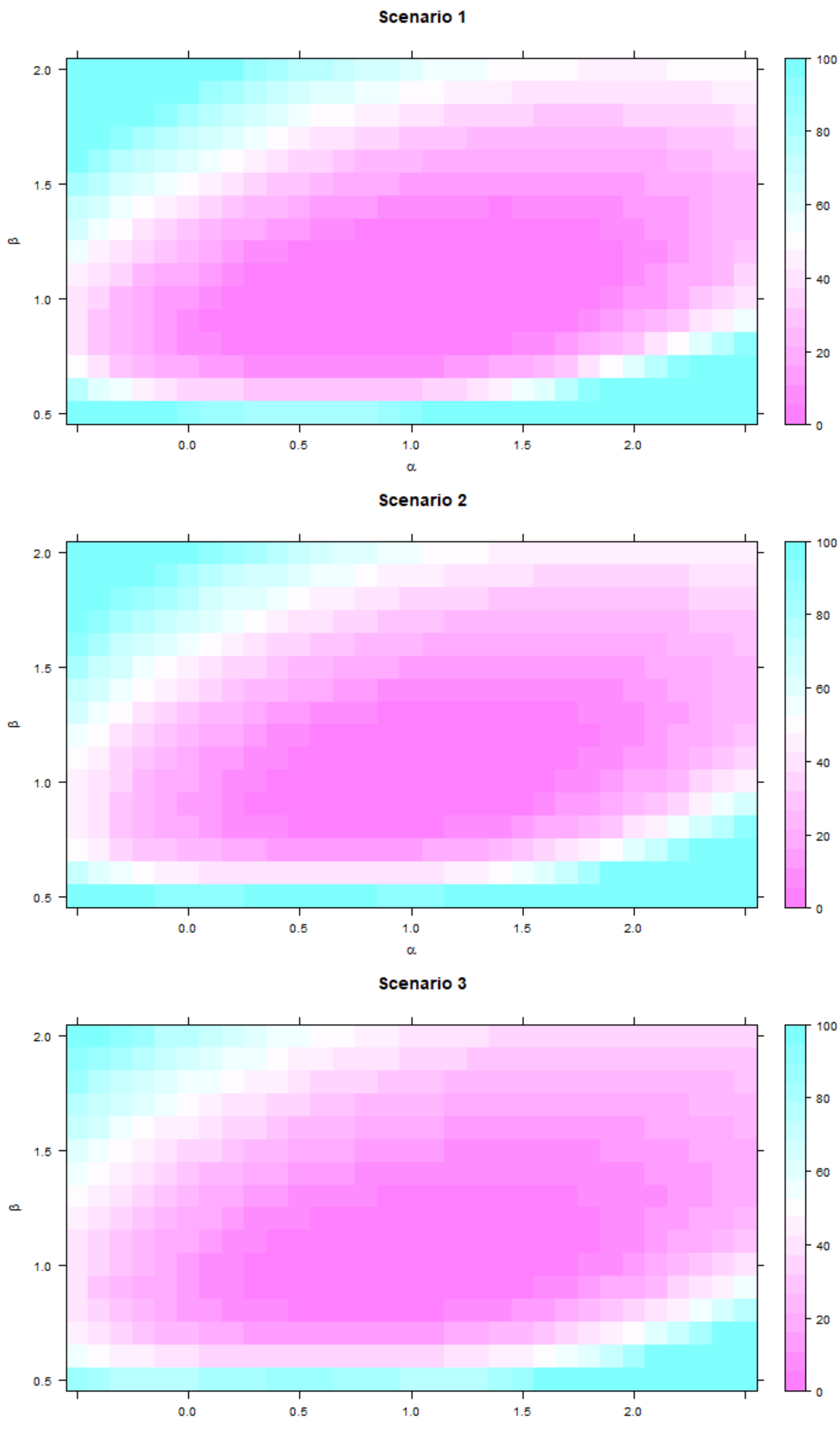

Figure 5.2: Gain in efficiency from one-stage to three-stage design, $N = 300$

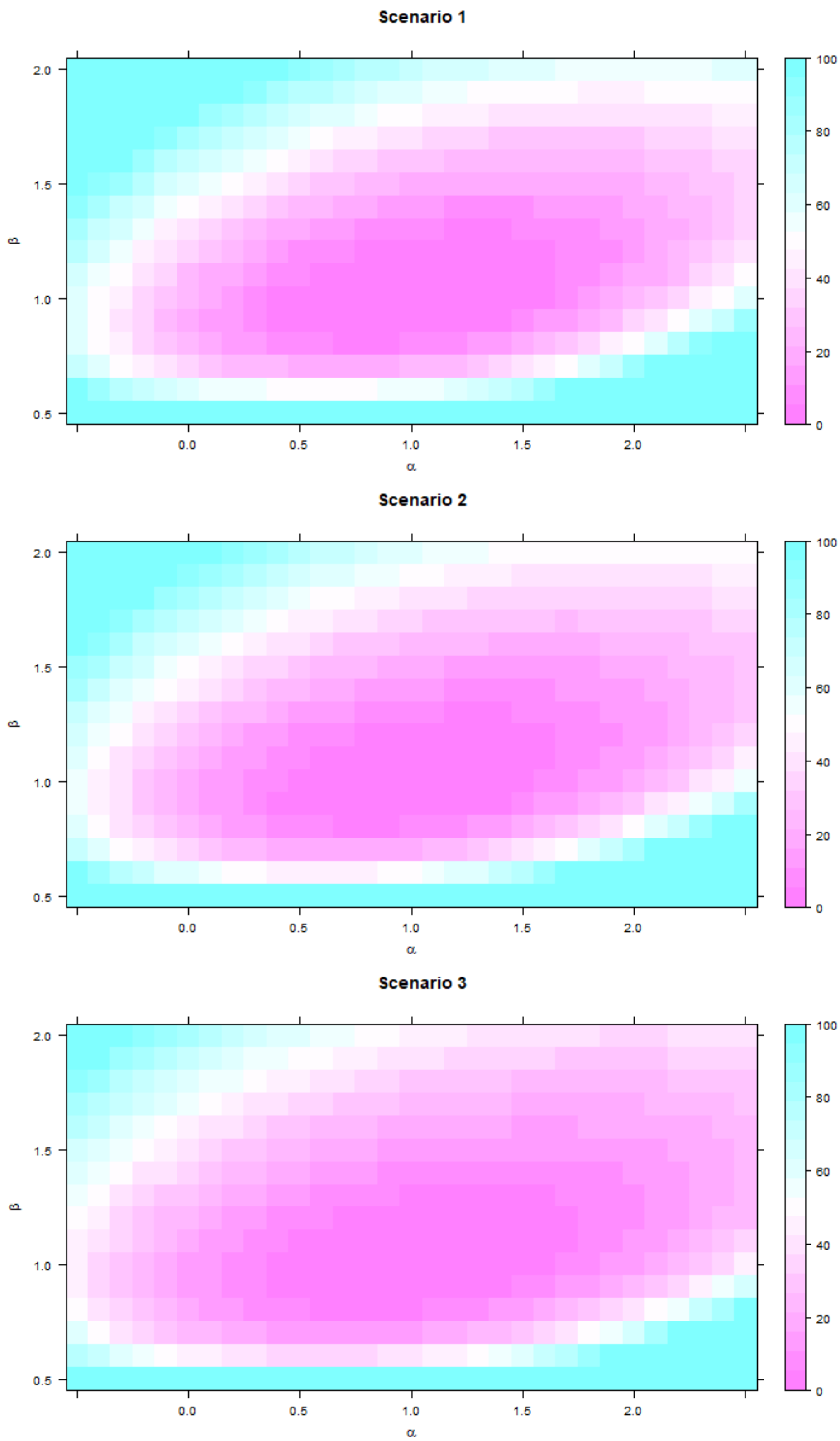

Figure 5.3: Gain in efficiency from one-stage to three-stage design, $N = 600$

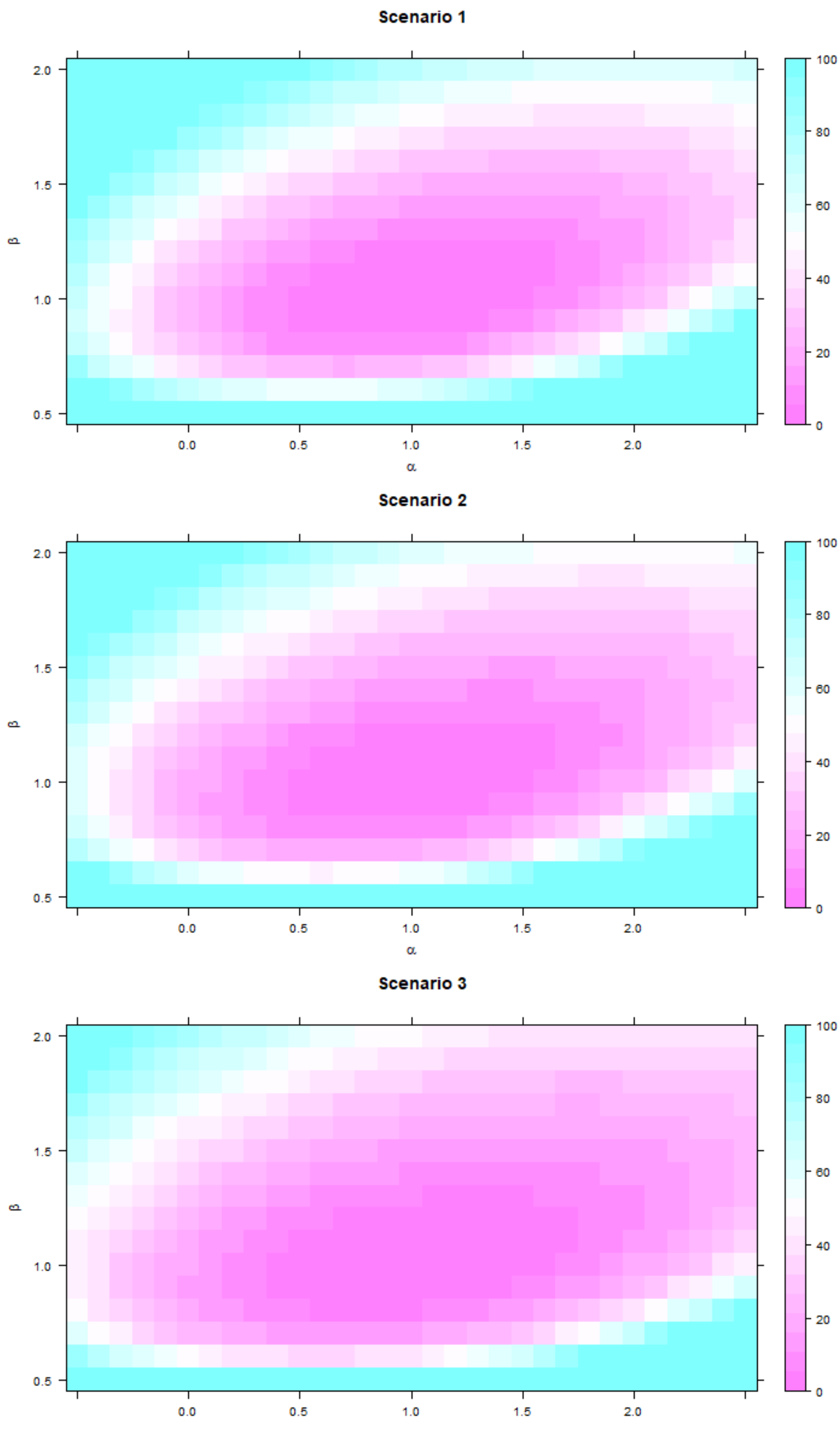

Figure 5.4: Gain in efficiency from one-stage to three-stage design, $N = 900$

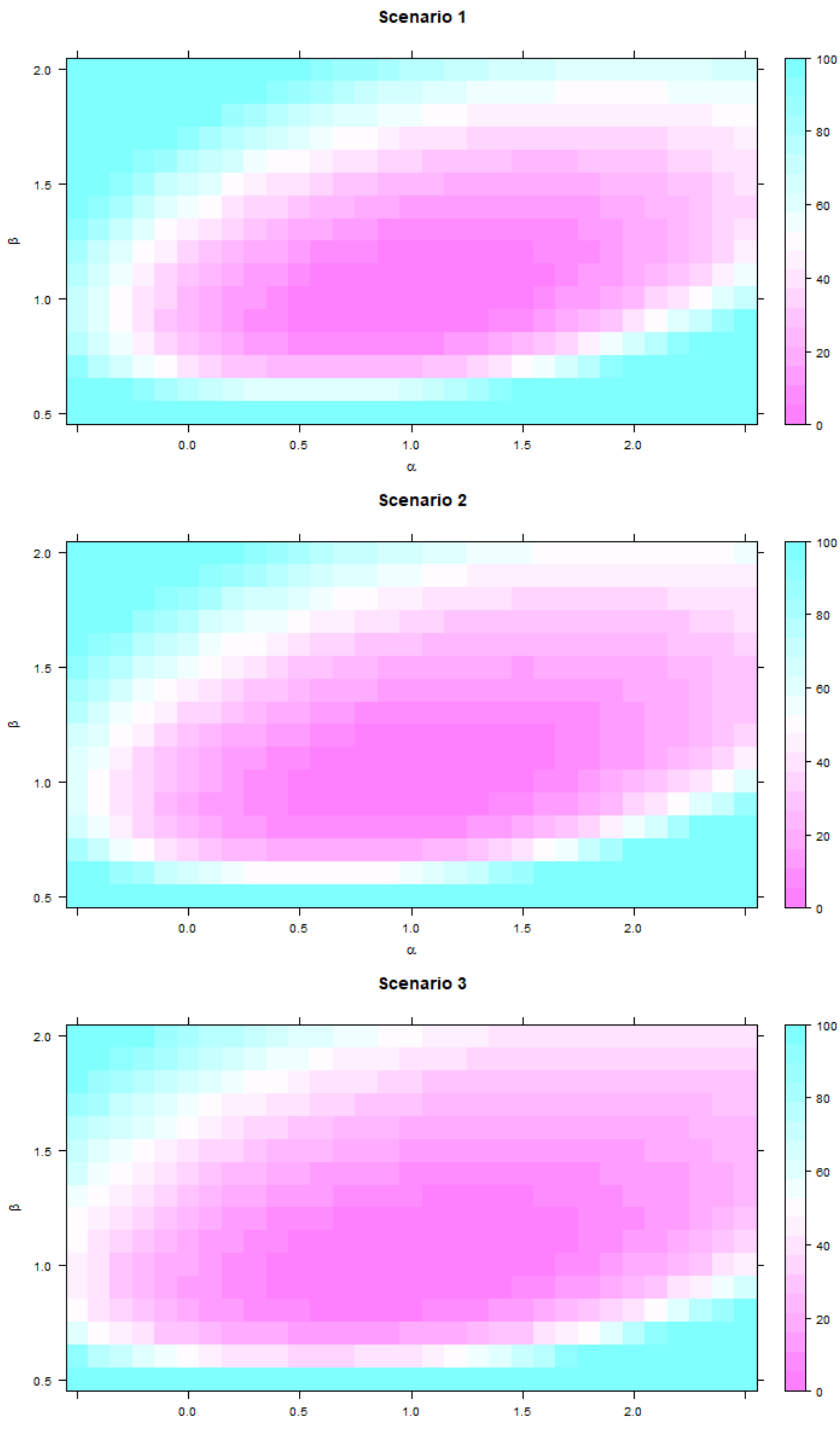

Figure 5.5: Gain in efficiency from one-stage to three-stage design, $N = 1200$

Figures 5.6 to 5.10 present the gain in efficiency from a two-stage design to a three-stage design in the form of a heat map for various sample sizes under each of the allocation scenarios presented in 5.1. The initial guess values of the unknown parameters α and β are depicted along the x and y axis respectively. Each (x, y) coordinate on the grid represents the gain in efficiency obtained when employing a three stage design compared to a reference two-stage design. For the sake of legibility of the plots, the peak gain in efficiency is set to 100% with any higher values also set to 100%. Similarly, any values less than 0% i.e. cases in which three-stage design is worse than the one-stage design is set to 0%. Table 5.2 presents the actual minimum, media, and maximum values of efficiency gained at various sample sizes under each scenario without aforementioned limits.

Inspecting the figures and the tables, we observe that the three-stage design results in a gain in efficiency over a one-stage design in most of the search space except for when the guess values are very close to the true values. When the guess values are far from the true values, the three stage design can result in a substantial increase in the efficiency.

		Gain in Efficiency From Two-Stage to Three-Stage		
		Min	Median	Max
N=150	Scenario 1	10.0%	19.9%	197.4%
	Scenario 2	2.4%	6.3%	132.0%
	Scenario 3	0.5%	2.0%	74.5%
N=300	Scenario 1	4.8%	11.0%	189.7%
	Scenario 2	1.1%	3.2%	126.1%
	Scenario 3	0.3%	1.0%	76.8%
N=600	Scenario 1	2.3%	5.4%	165.3%
	Scenario 2	0.6%	1.5%	120.9%
	Scenario 3	0.1%	0.4%	68.1%
N=900	Scenario 1	1.5%	3.5%	163.2%
	Scenario 2	0.4%	0.9%	107.5%
	Scenario 3	0.1%	0.3%	57.1%
N=1200	Scenario 1	1.1%	2.5%	154.7%
	Scenario 2	0.3%	0.7%	95.7%
	Scenario 3	0.1%	0.2%	50.0%

Table 5.3: Gain in efficiency from two-stage to three-stage design

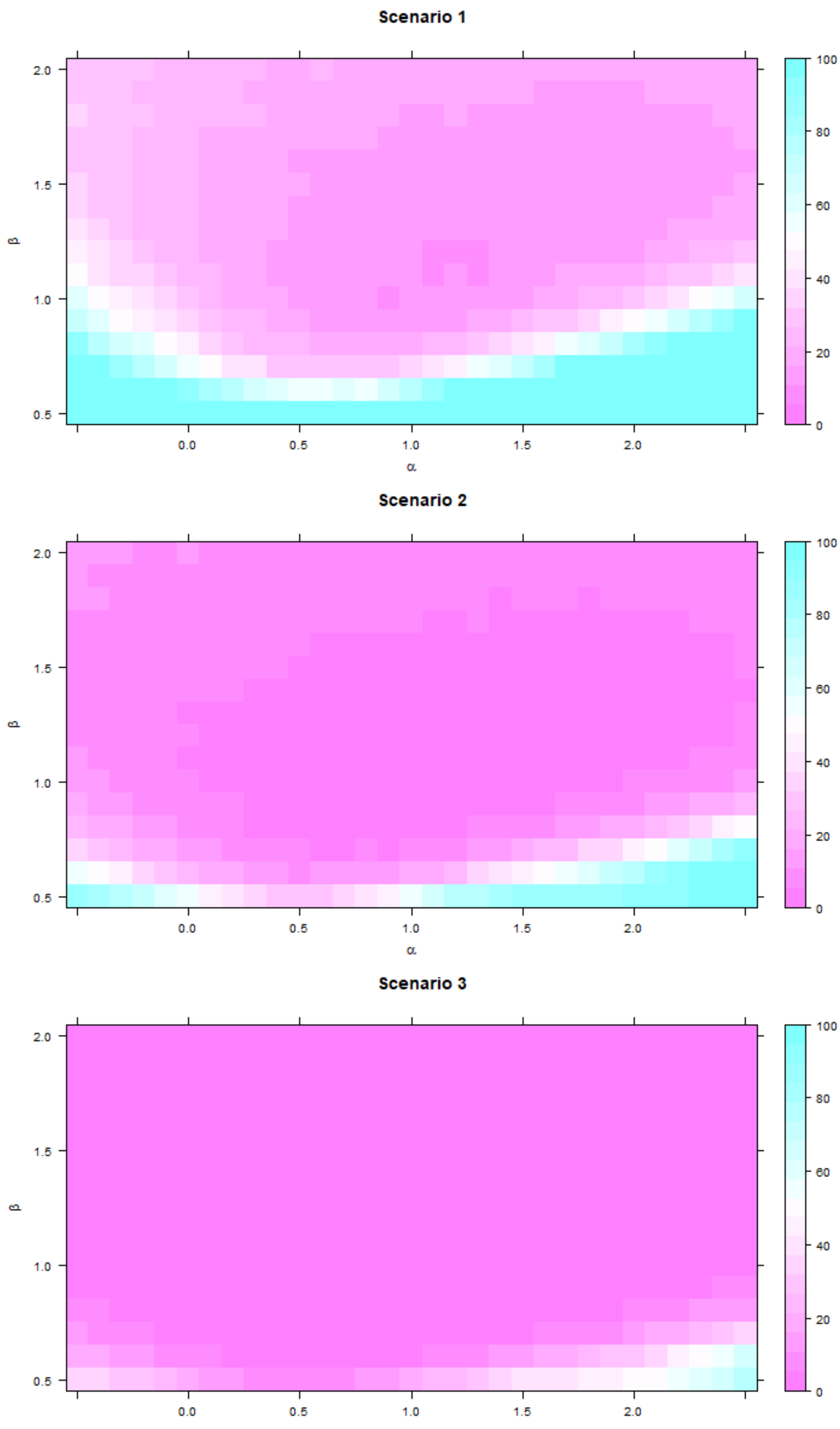

Figure 5.6: Gain in efficiency from two-stage to three-stage design, $N = 150$

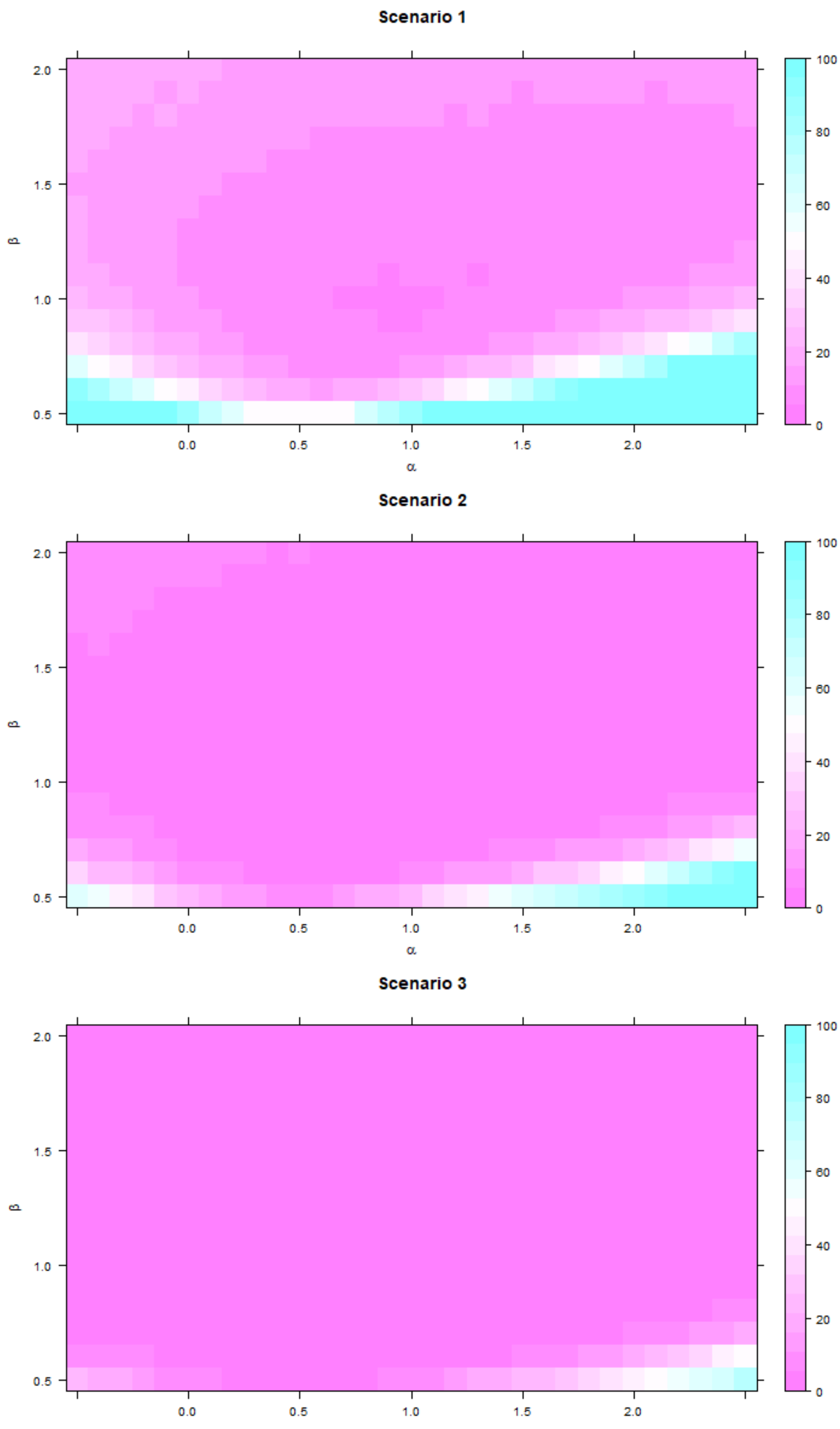

Figure 5.7: Gain in efficiency from two-stage to three-stage design, $N = 300$

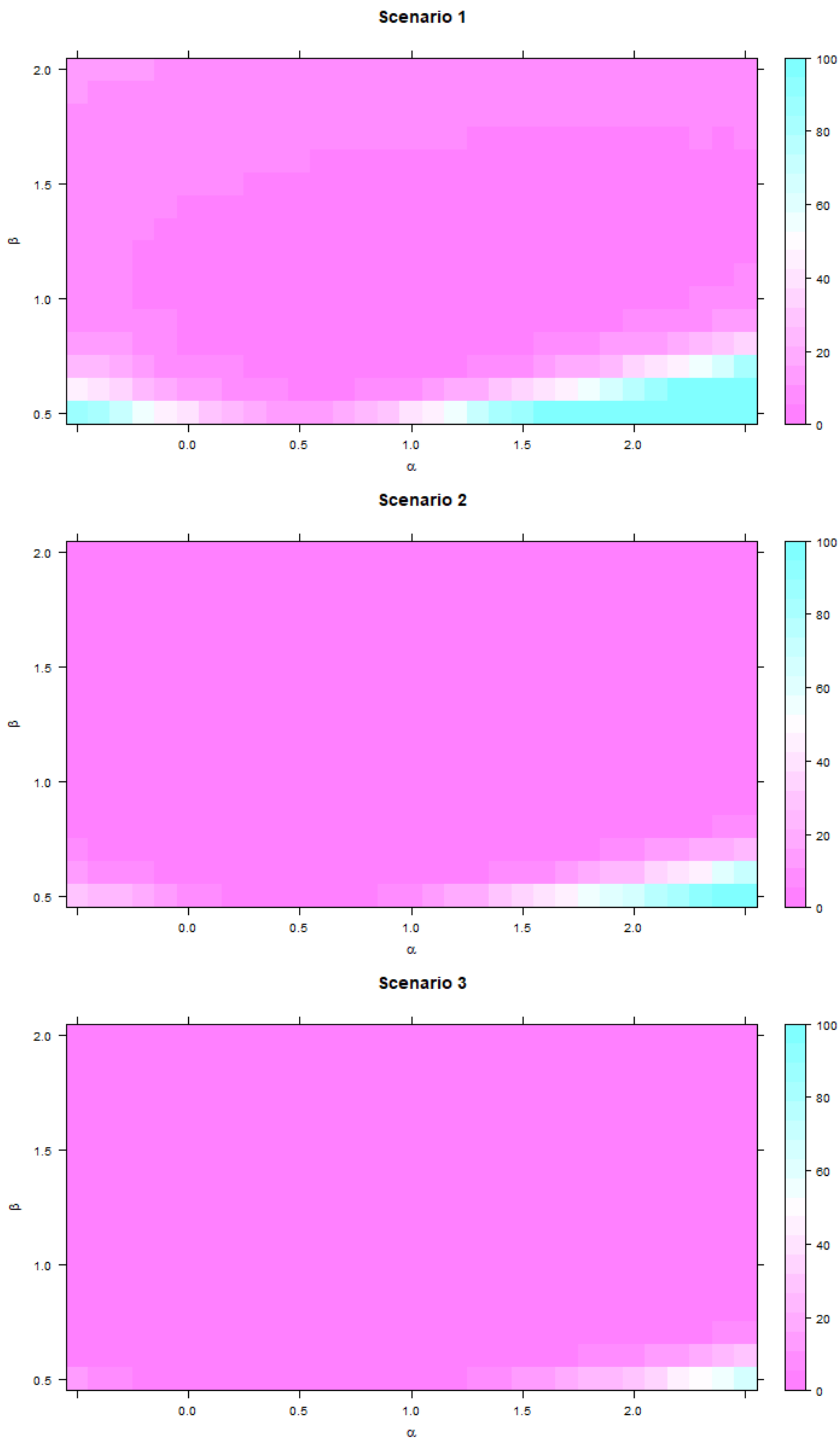

Figure 5.8: Gain in efficiency from two-stage to three-stage design, $N = 600$

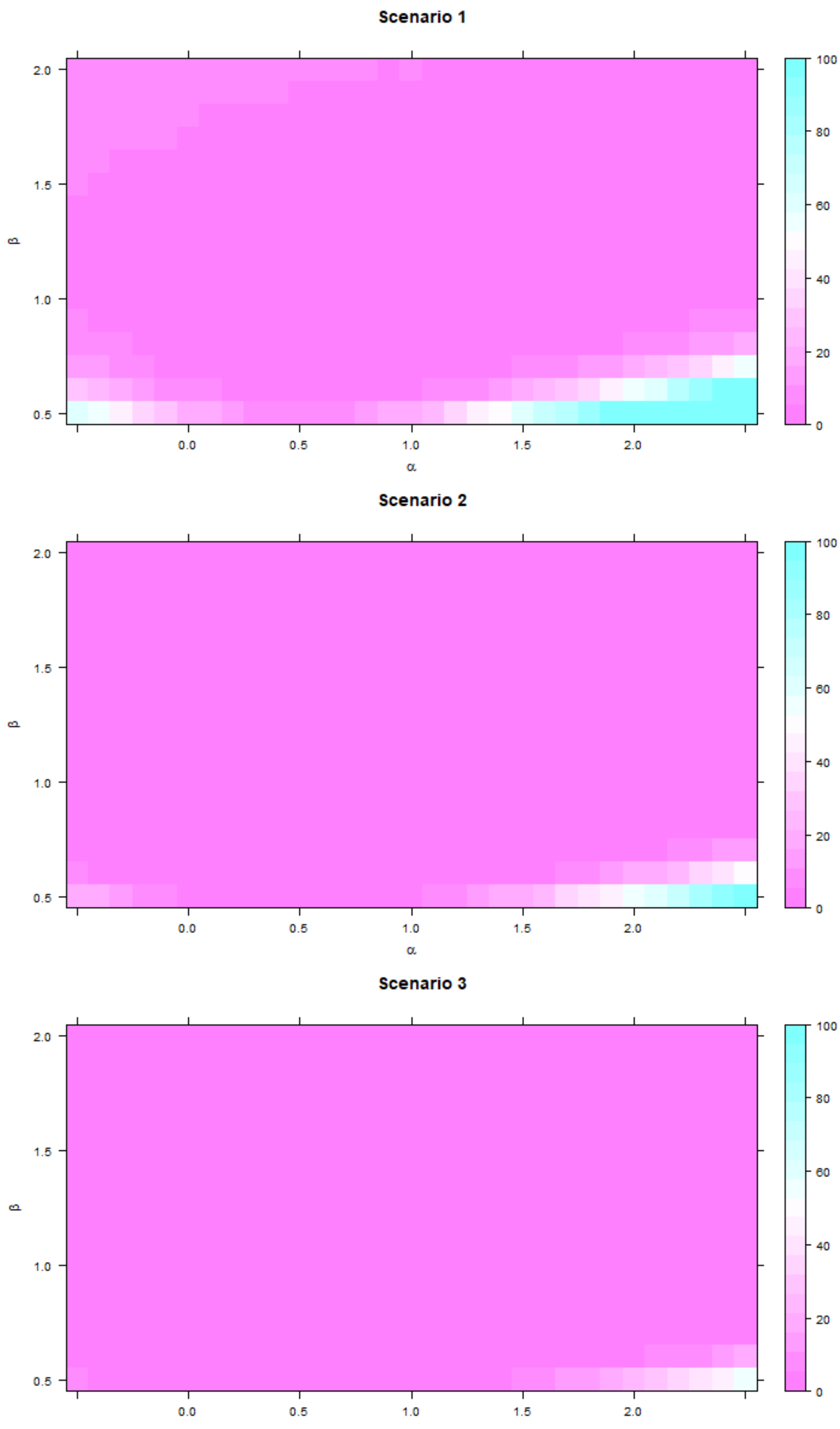

Figure 5.9: Gain in efficiency from two-stage to three-stage design, $N = 900$

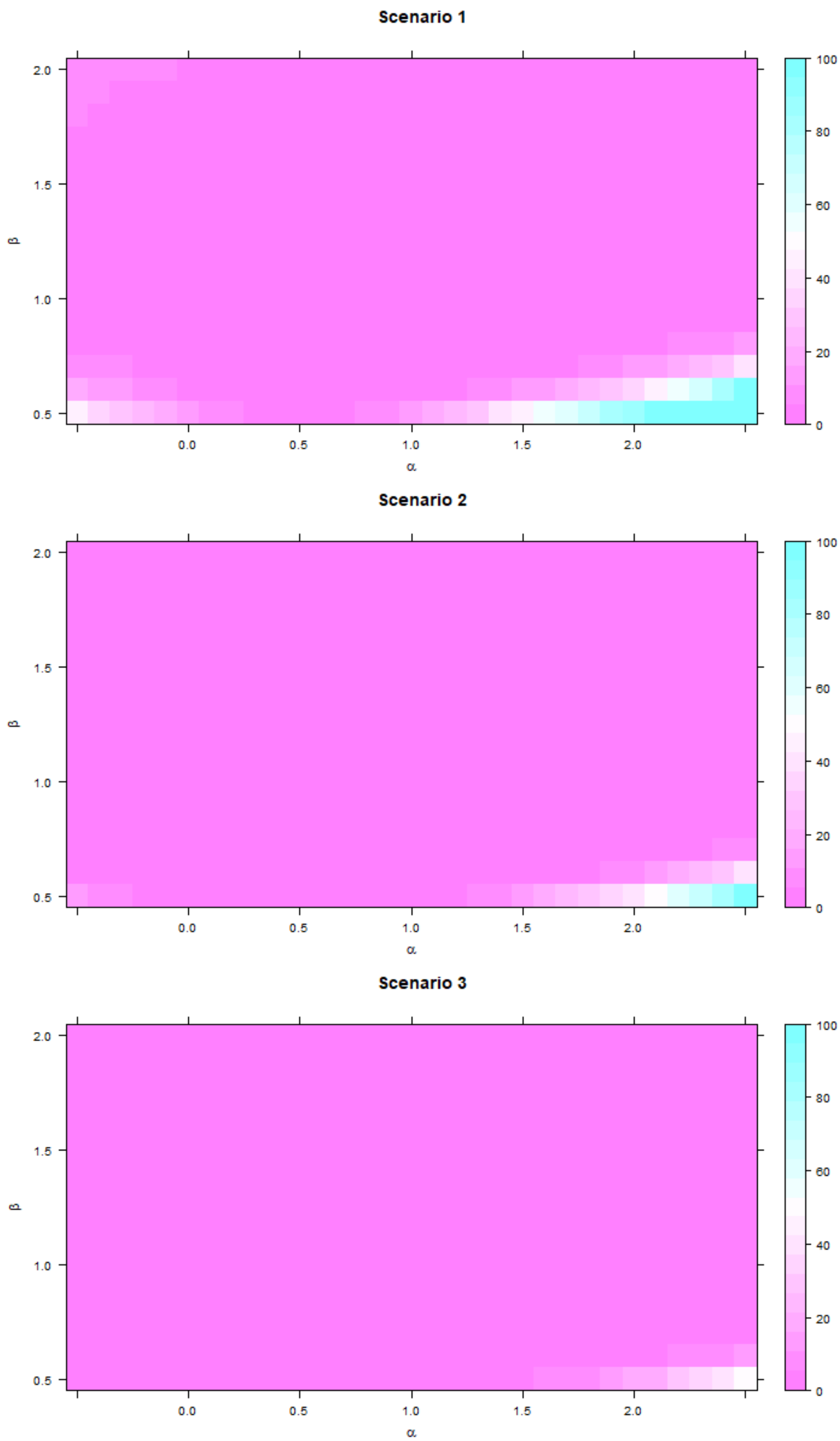

Figure 5.10: Gain in efficiency from two-stage to three-stage design, $N = 1200$

5.4 Conclusions & Recommendations

Although the D-optimal criterion is independent of sample size, we see that there is an observable impact of total sample size and the proportion allocated to each stage. There are two main reasons for this:

1. The three stage optimality criterion is a weighted average of the D-optimality criterion at each stage and therefore, the proportion of sample allocated to each stage creates some dependency on the sample size.

2. The iterative approach to refining the estimates of the unknown parameters α and β mitigates the impact of poor initial guess values

We observe that the three stage D-optimal design performs at least as well as the two-stage design on average when the guess values are close to the true values. The further the guess values are from the true values, the greater the gain in efficiency.

When the sample size is small ($N = 150$ to $N = 300$) Scenario 1, with a small proportion allocated to initial stage and progressively increasing the proportion of data allocated at each stage, results in the largest gain across the search space. The median efficiency achieved at these sample sizes is 10-20 % more than the corresponding two-stage design. At larger sample sizes and ascending dose allocation works best.

Chapter 6

Practitioner Recommendations

In this thesis we have considered practical issues concerned with the implementation of optimal designs namely the issues related to the sample size and unknown parameter estimates for a two parameter logistic regression model on which the optimal design depends. Our main recommendations arising from the specific aims are as follows:

1. For finite samples significant improvement can be achieved by solving the A-optimality problem numerically within a restricted space.

2. A two-stage A-optimal design can result in increased efficiency under specific conditions. When the total sample size available for the experiment is small, allocating 70% of the sample to the initial stage of two-stage finite sample design can result in improvements of $60-70\%$ compared to a single stage finite sample design.

3. For large sample D-optimality designs, a three stage design performs at least as well as a two-stage design, on average. For small sample sizes, an approach of increasing the proportion of the total sample size allocated to progressive stages results in a $10-20\%$ median gain at $N = 100$ & $N = 150$. For large samples the increase is $2-5\%$.

The methods employed in this work can be employed to evaluate logistic models with more parameters e.g. a quadratic logistic model, and to other underlying models such as a Poisson regression model. Future work may also consider study specific considerations and restrictions such as limits on dose values due to safety considerations.

Appendix A

Additional Aim 3 Results

		Gain in Efficiency (%)		
		Min	Median	Max
N=150	Scenario 1	-20.5	12.4	4919.3
	Scenario 2	-7.4	21.4	4433.4
	Scenario 3	-4.0	21.1	2622.5
N=300	Scenario 1	-7.6	33.0	6701.8
	Scenario 2	-3.5	36.5	4868.4
	Scenario 3	-1.2	29.0	2829.3
N=600	Scenario 1	-3.5	48.2	7287.3
	Scenario 2	-1.8	42.3	5206.1
	Scenario 3	-1.0	32.4	3001.2
N=900	Scenario 1	-2.3	52.4	7555.7
	Scenario 2	-1.1	43.9	5379.3
	Scenario 3	-0.7	33.4	3121.3
N=1200	Scenario 1	-1.7	54.2	7724.6
	Scenario 2	-0.9	44.6	5495.1
	Scenario 3	-0.5	33.8	3235.3

Table A.1: Gain in efficiency from one-stage to three-stage, $(\alpha_t = 2, \beta_t = 1)$

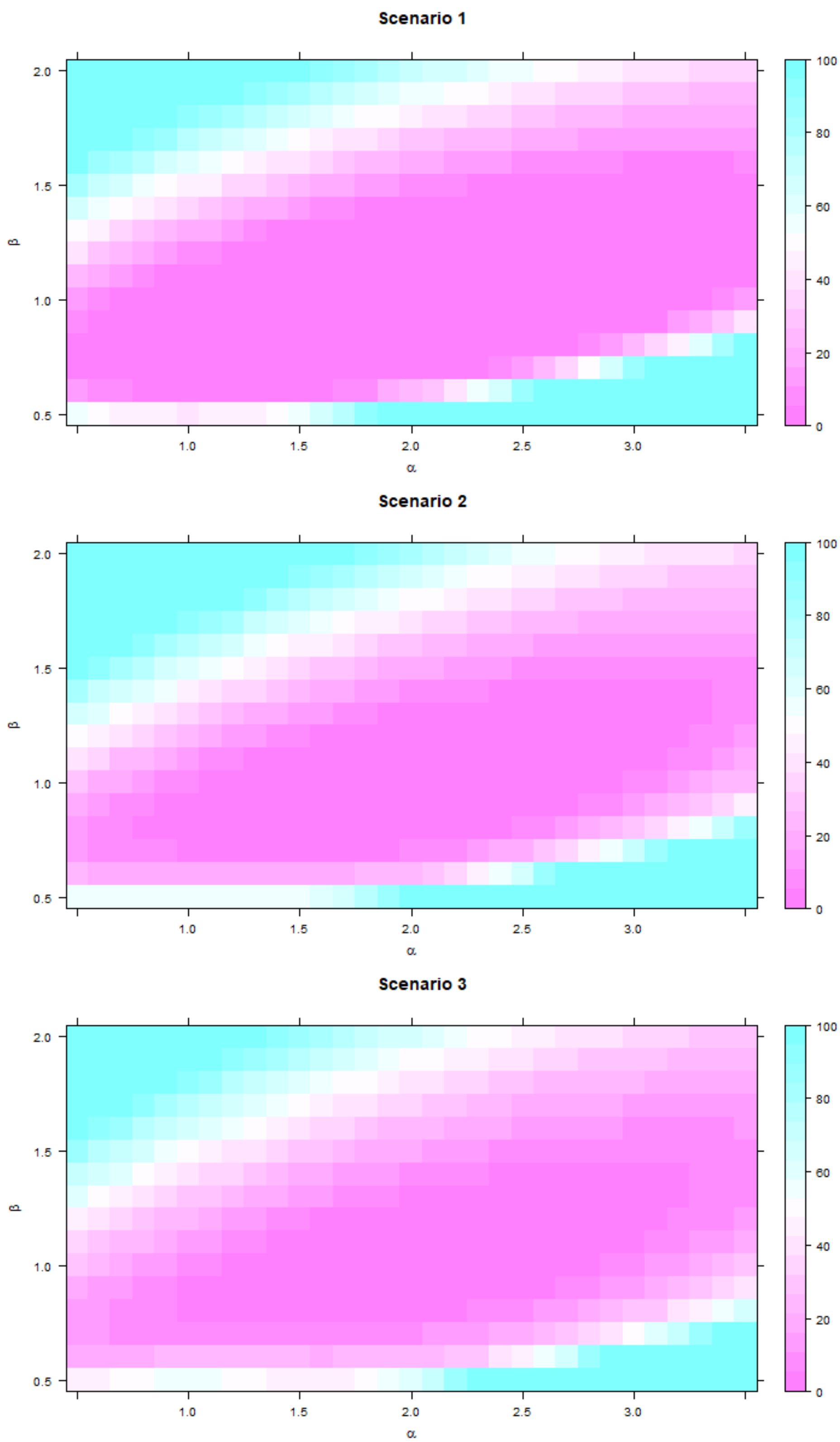

Figure A.1: Gain in efficiency from one-stage to three-stage design

$N = 150$, $(\alpha_t = 2, \beta_t = 1)$

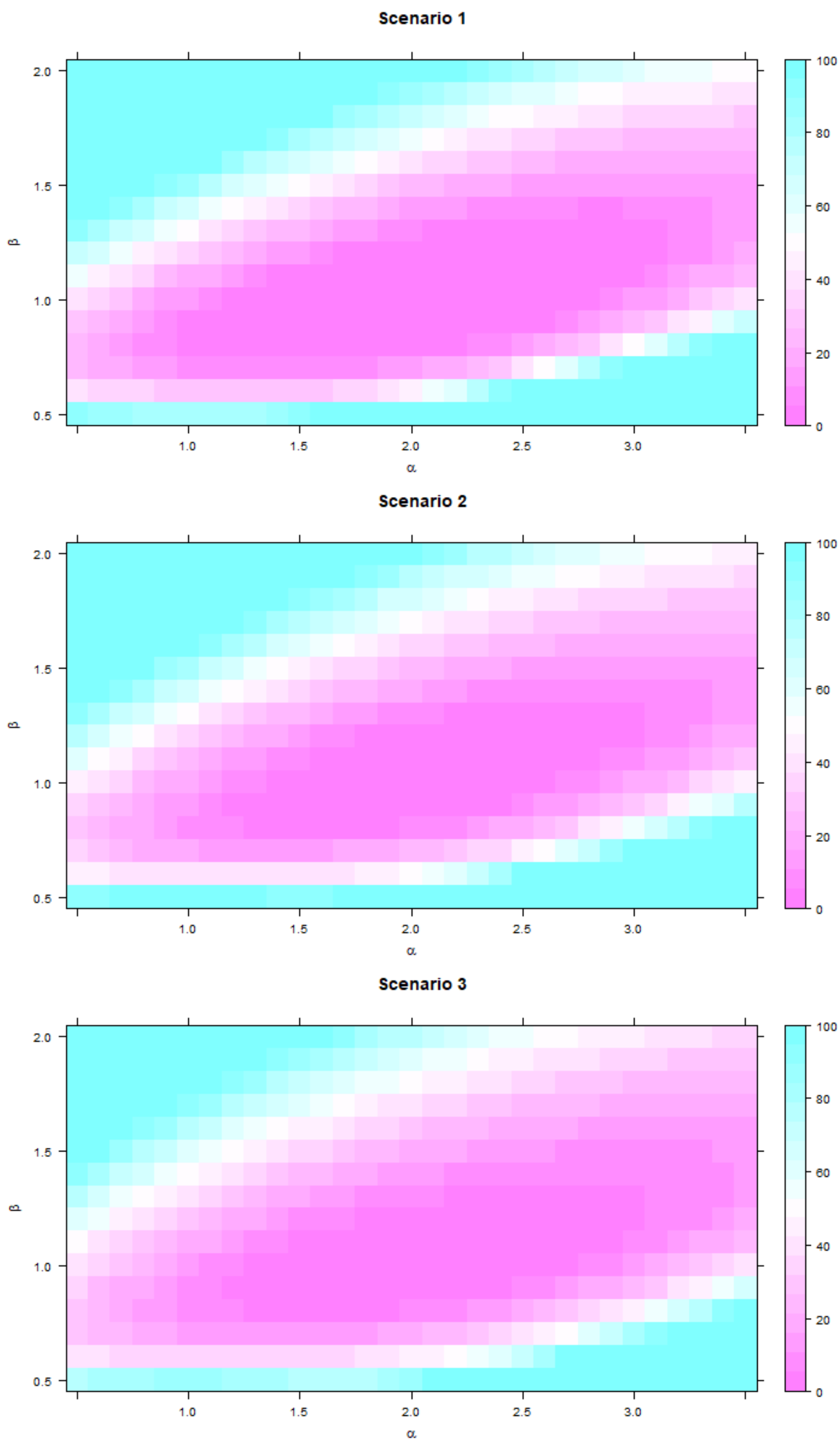

Figure A.2: Gain in efficiency from one-stage to three-stage design

$N = 300$, $(\alpha_t = 2, \beta_t = 1)$

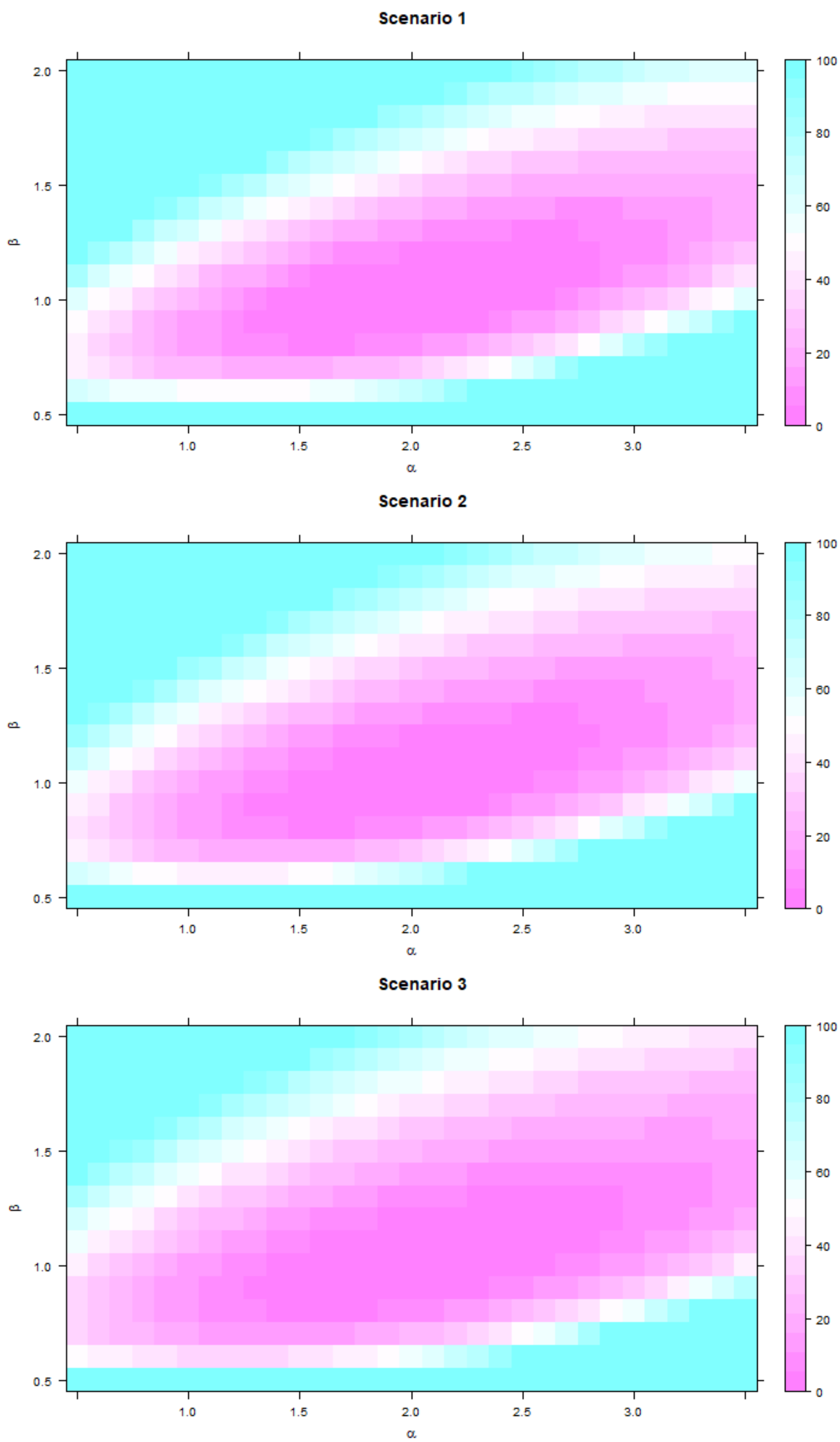

Figure A.3: Gain in efficiency from one-stage to three-stage design

$N = 600$, $(\alpha_t = 2, \beta_t = 1)$

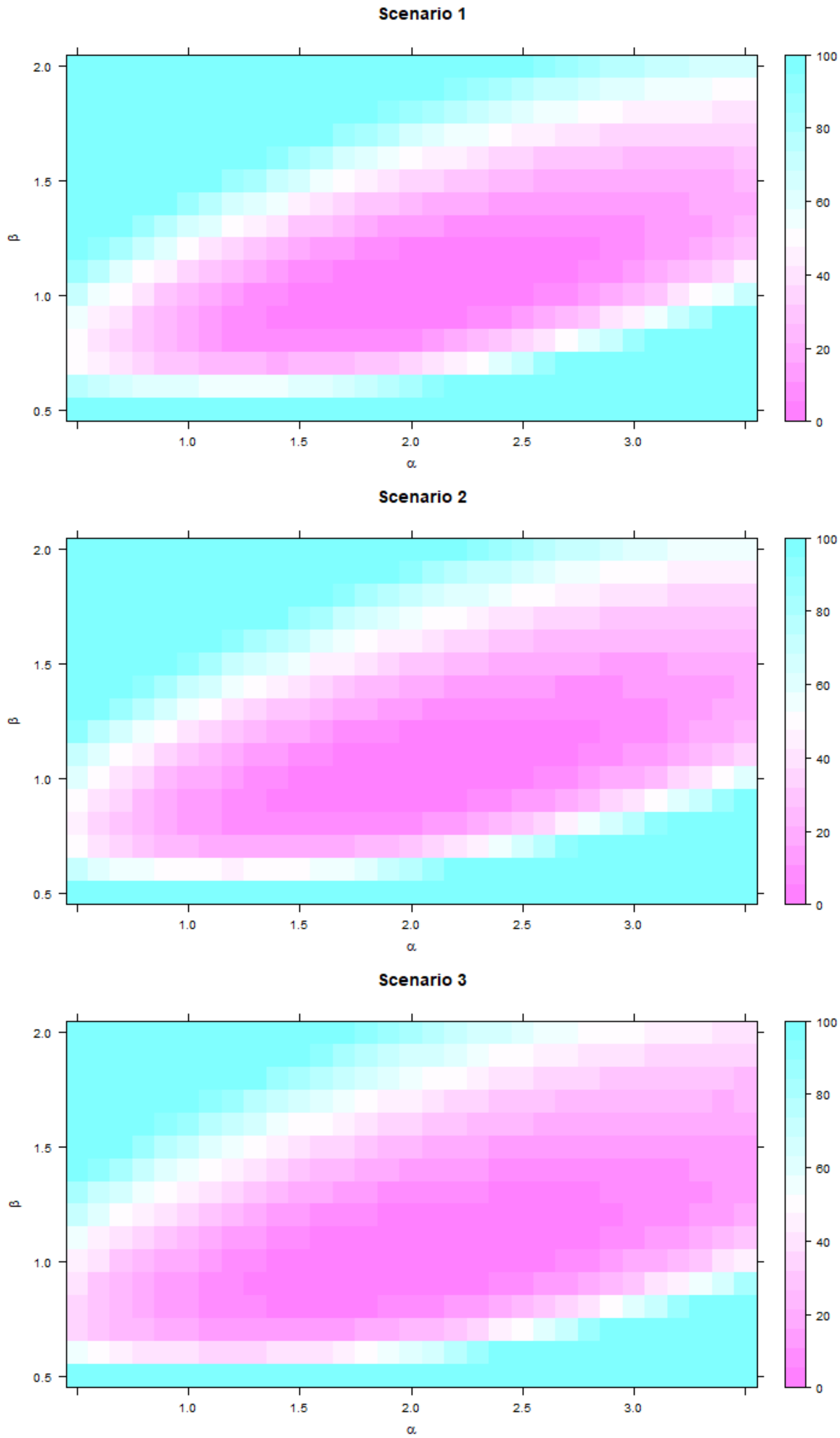

Figure A.4: Gain in efficiency from one-stage to three-stage design

$N = 900$, $(\alpha_t = 2, \beta_t = 1)$

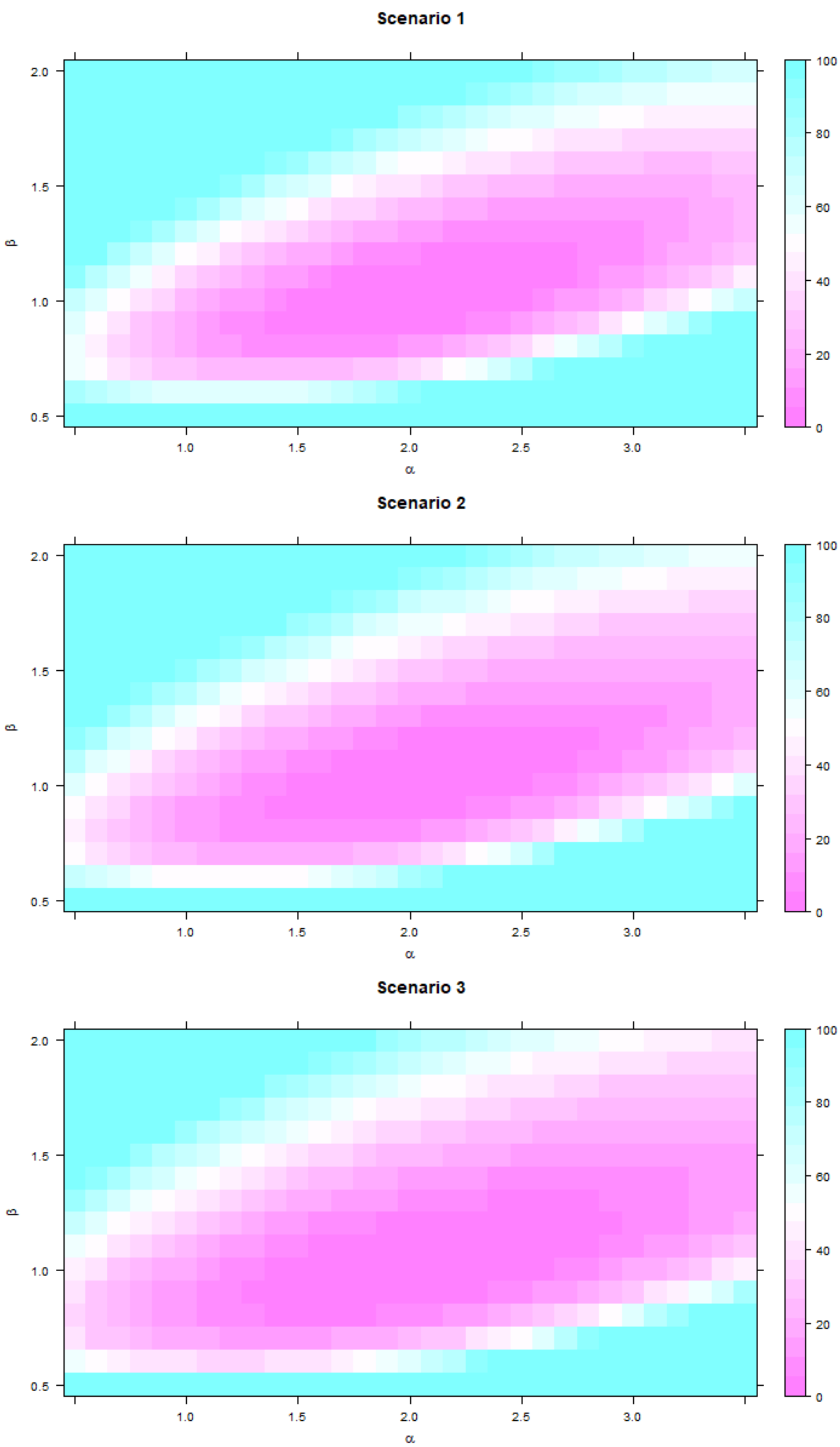

Figure A.5: Gain in efficiency from one-stage to three-stage design

$N = 1200$, $(\alpha_t = 2, \beta_t = 1)$

		Gain in Efficiency (%)		
		Min	Median	Max
N=150	Scenario 1	10.3	24.8	250.3
	Scenario 2	2.2	7.2	202.4
	Scenario 3	0.5	2.3	134.9
N=300	Scenario 1	4.8	12.6	280.3
	Scenario 2	1.1	3.6	232.5
	Scenario 3	0.3	1.1	159.4
N=600	Scenario 1	2.2	6.2	299.9
	Scenario 2	0.6	1.6	261.8
	Scenario 3	0.1	0.5	168.2
N=900	Scenario 1	1.4	3.9	322.3
	Scenario 2	0.4	1.0	264.8
	Scenario 3	0.1	0.3	162.6
N=1200	Scenario 1	1.1	2.7	332.1
	Scenario 2	0.3	0.8	256.9
	Scenario 3	0.1	0.2	151.0

Table A.2: Gain in efficiency from two-Stage to three-Stage, $(\alpha_t = 2, \beta_t = 1)$

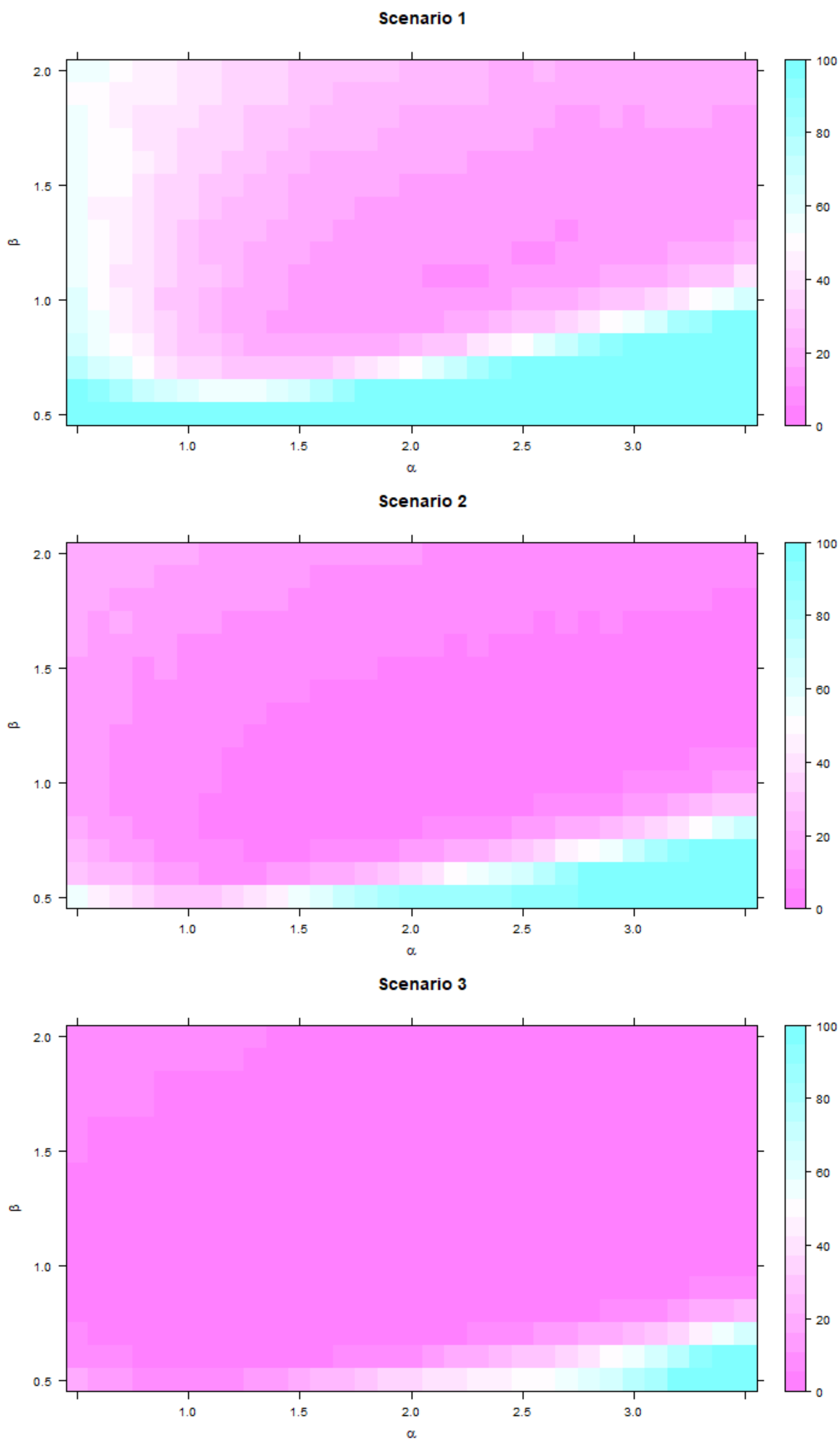

Figure A.6: Gain in efficiency from one-stage to three-stage design

$N = 150$, $(\alpha_t = 2, \beta_t = 1)$

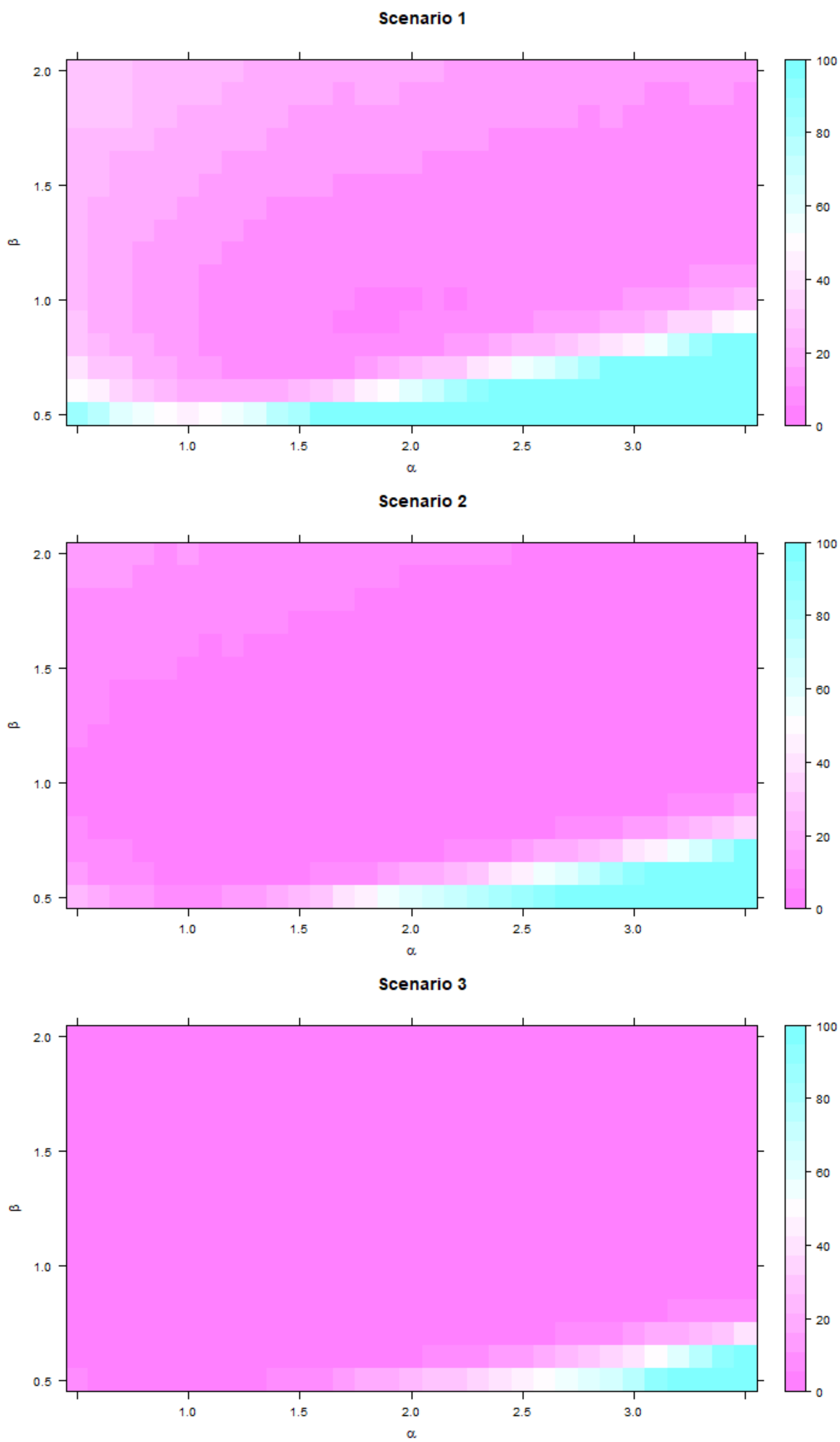

Figure A.7: Gain in efficiency from one-stage to three-stage design

$N = 300$, $(\alpha_t = 2, \beta_t = 1)$

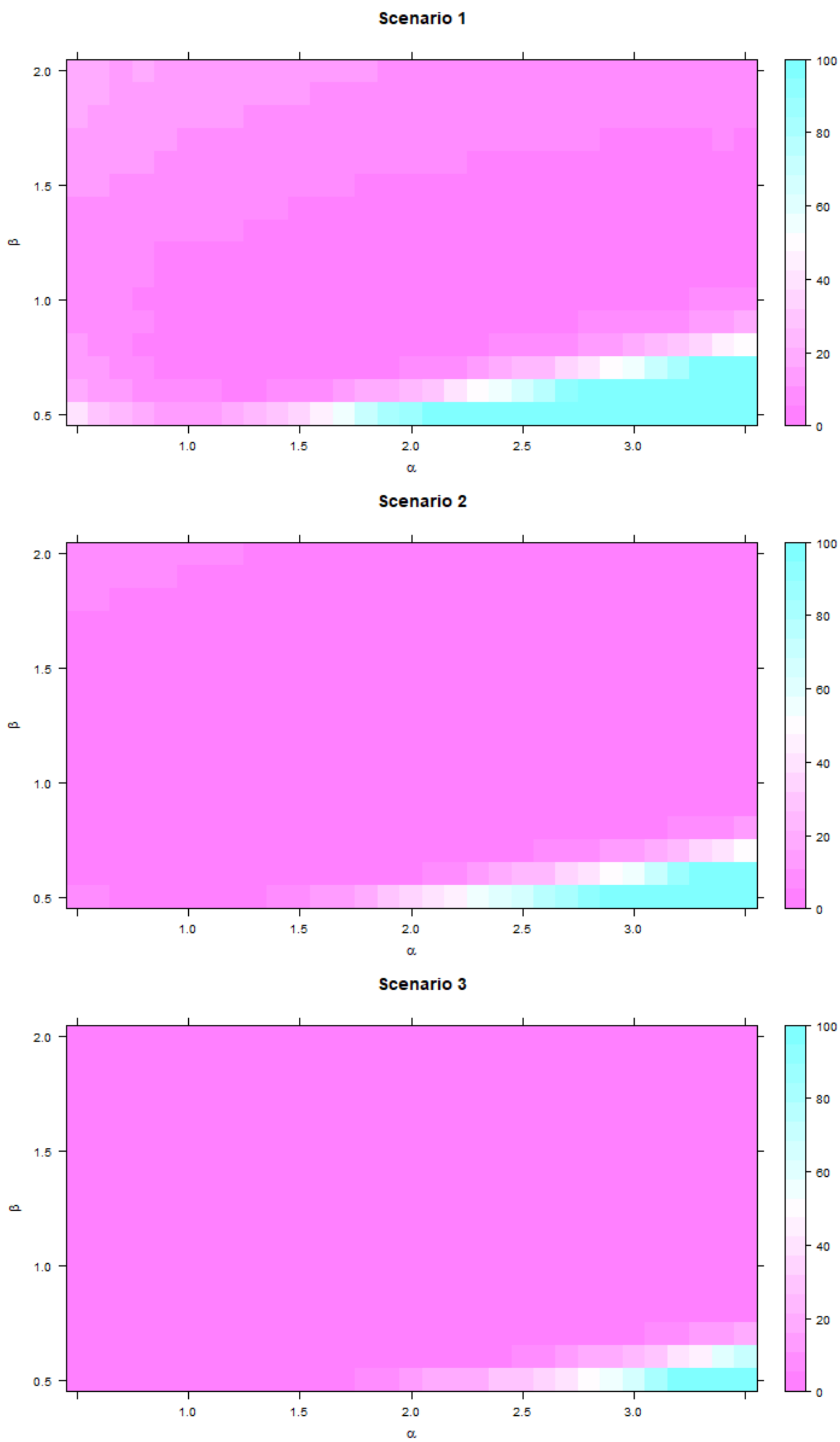

Figure A.8: Gain in efficiency from one-stage to three-stage design

$N = 600$, $(\alpha_t = 2, \beta_t = 1)$

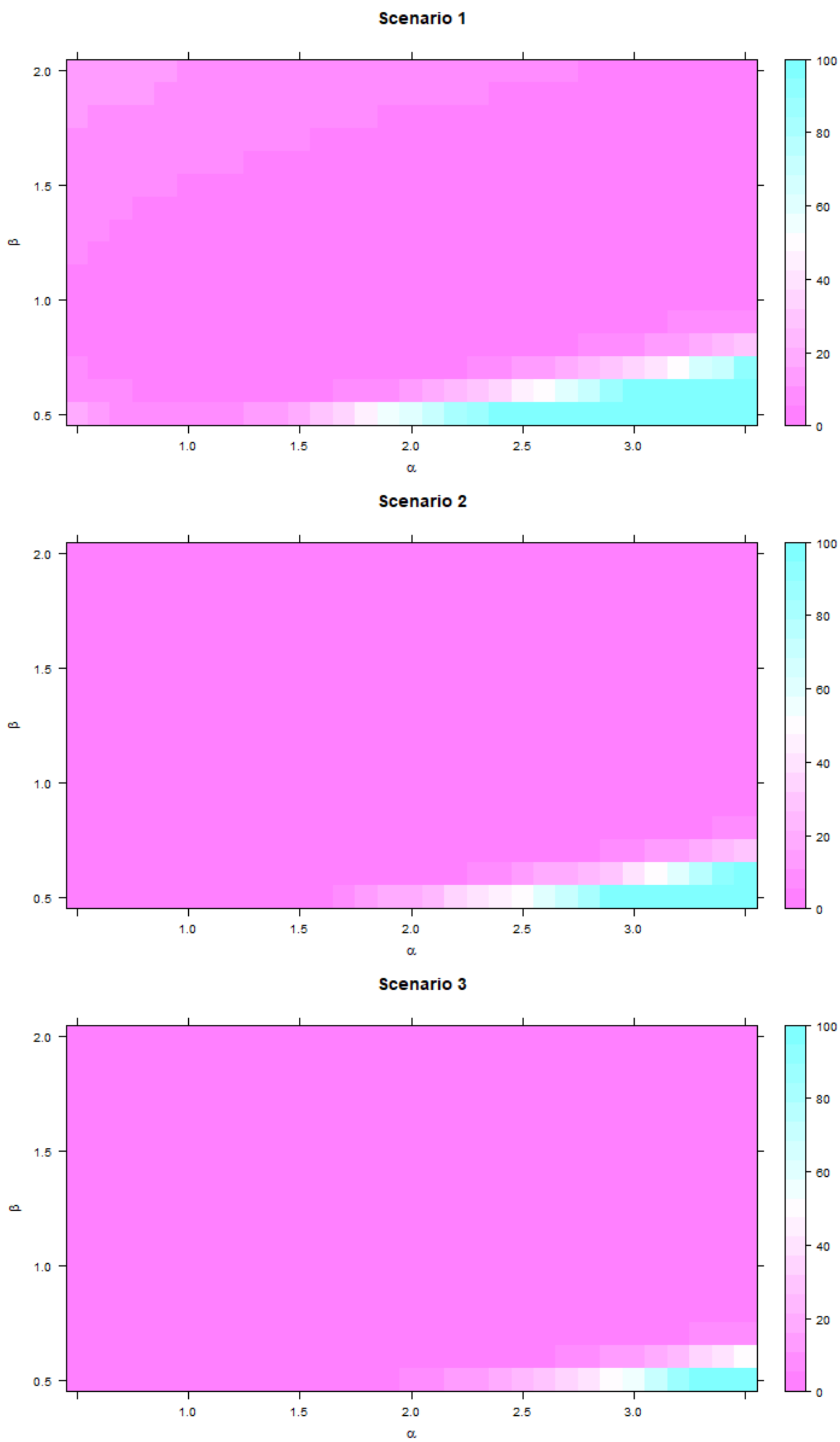

Figure A.9: Gain in efficiency from one-stage to three-stage design

$N = 900$, $(\alpha_t = 2, \beta_t = 1)$

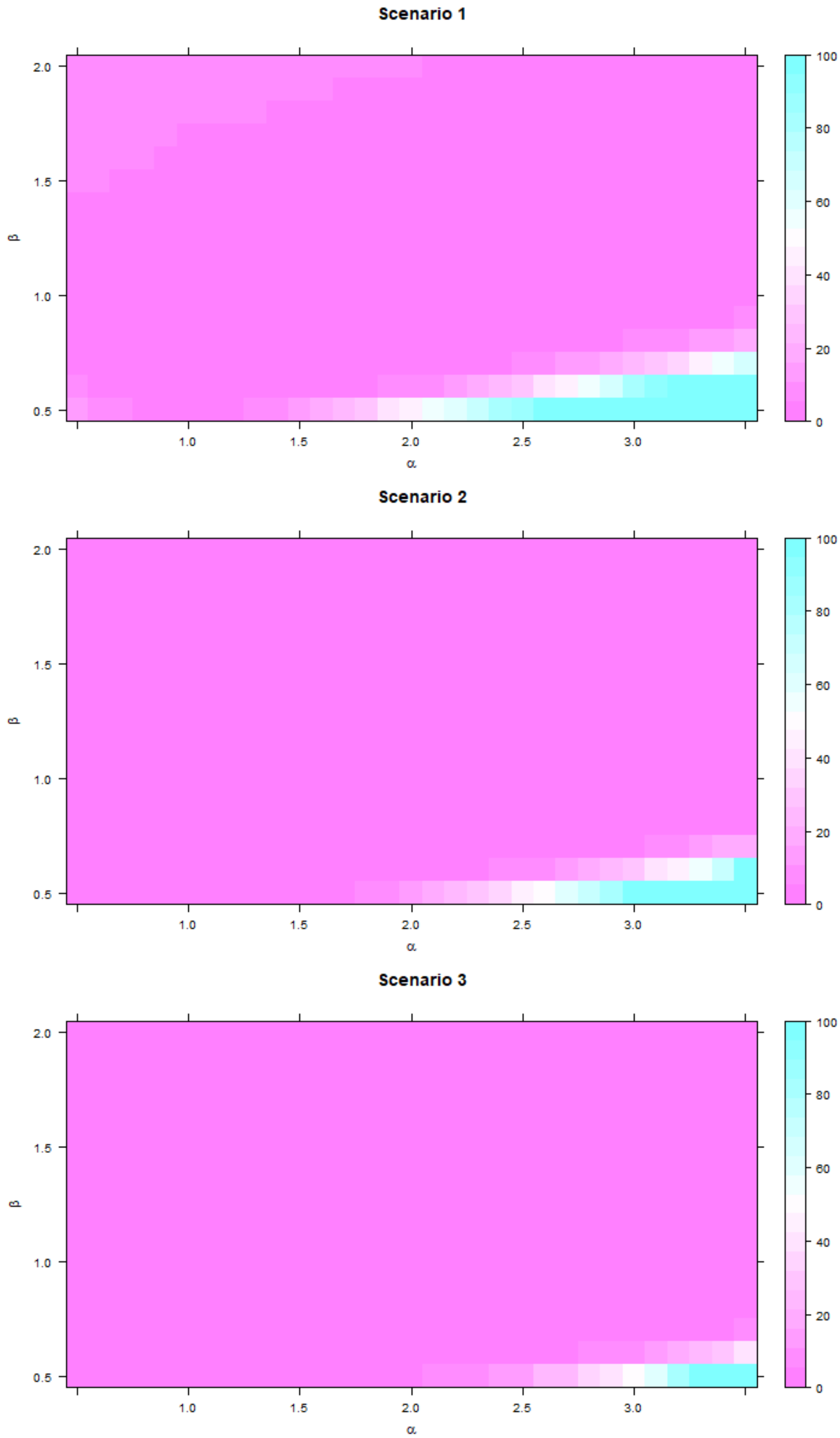

Figure A.10: Gain in efficiency from one-stage to three-stage design

$N = 1200$, $(\alpha_t = 2, \beta_t = 1)$

		Gain in Efficiency (%)		
		Min	Median	Max
N=150	Scenario 1	-20.6	8.2	1317.0
	Scenario 2	-7.4	18.7	996.1
	Scenario 3	-4.0	18.0	613.1
N=300	Scenario 1	-7.6	28.9	1460.2
	Scenario 2	-3.5	29.8	1080.1
	Scenario 3	-2.0	23.8	662.3
N=600	Scenario 1	-3.5	39.2	1564.1
	Scenario 2	-1.7	34.2	1155.4
	Scenario 3	-1.0	26.4	730.0
N=900	Scenario 1	-2.3	42.4	1618.3
	Scenario 2	-1.1	35.7	1207.9
	Scenario 3	-0.7	27.2	783.8
N=1200	Scenario 1	-1.7	43.8	1650.5
	Scenario 2	-0.9	36.4	1250.7
	Scenario 3	-0.5	27.6	827.7

Table A.3: Gain in efficiency from one-stage to three-stage, ($\alpha_t = -1, \beta_t = 1$)

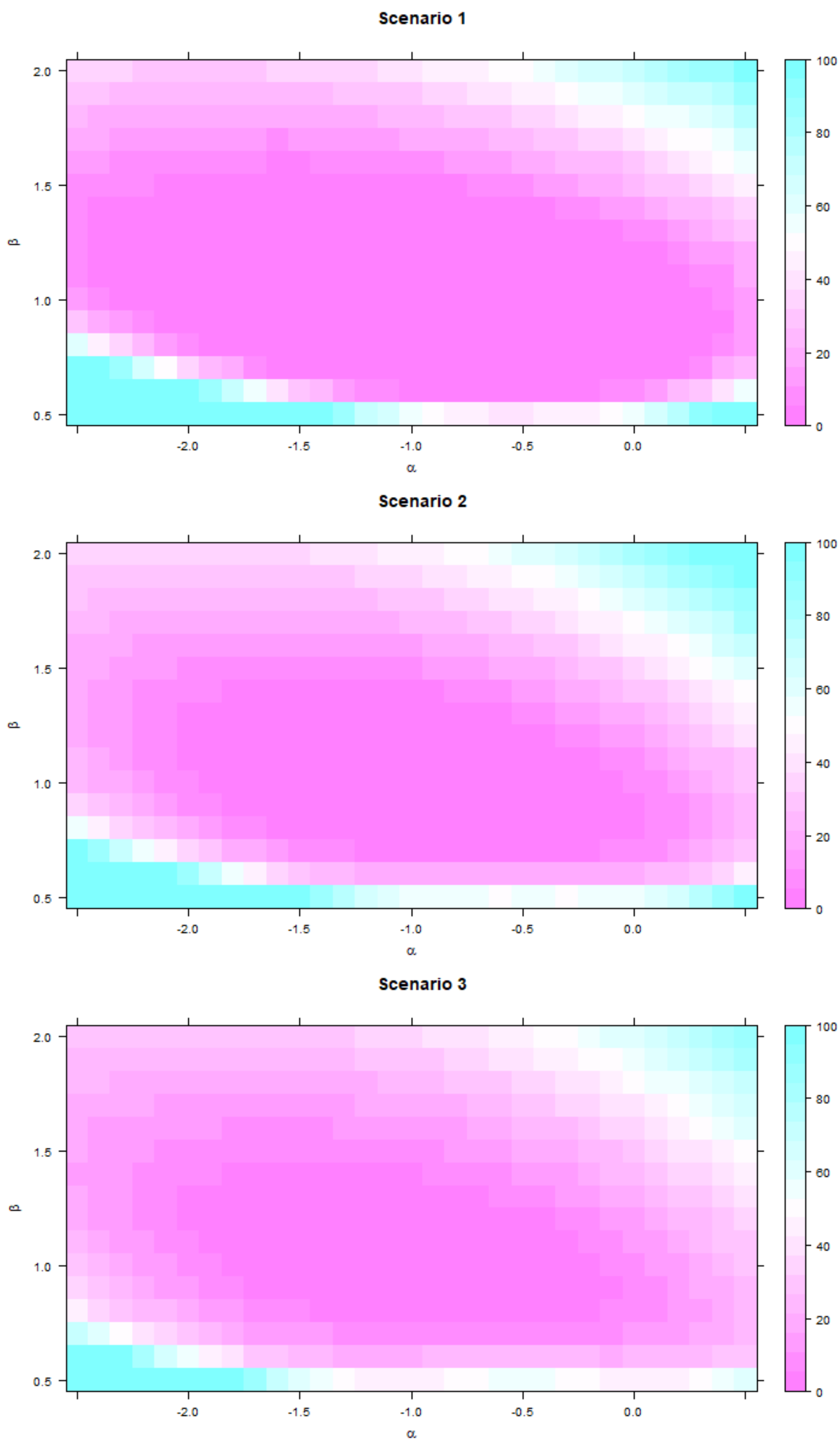

Figure A.11: Gain in efficiency from one-stage to three-stage design

$N = 150$, $(\alpha_t = -1, \beta_t = 1)$

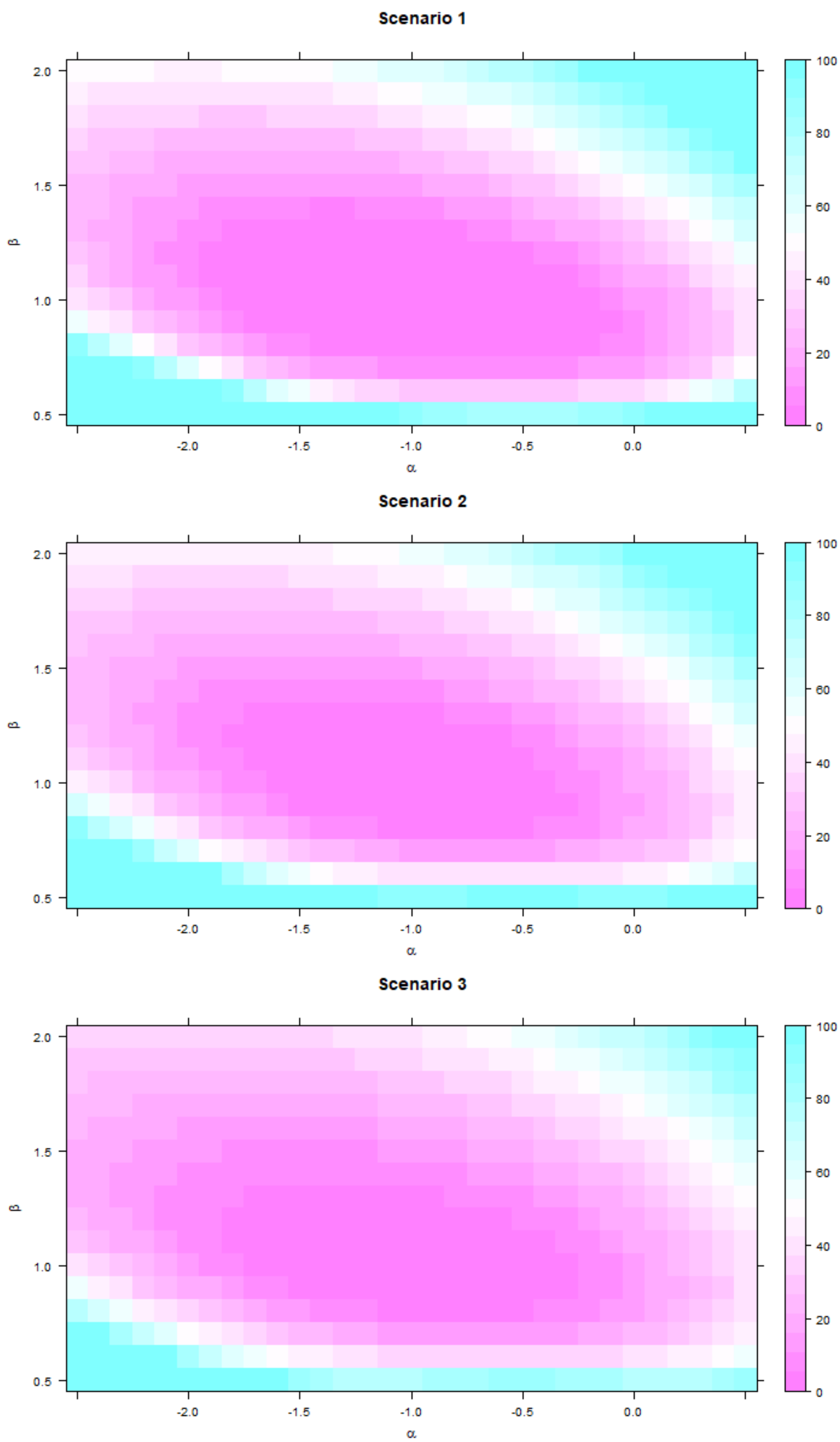

Figure A.12: Gain in efficiency from one-stage to three-stage design

$N = 300$, $(\alpha_t = -1, \beta_t = 1)$

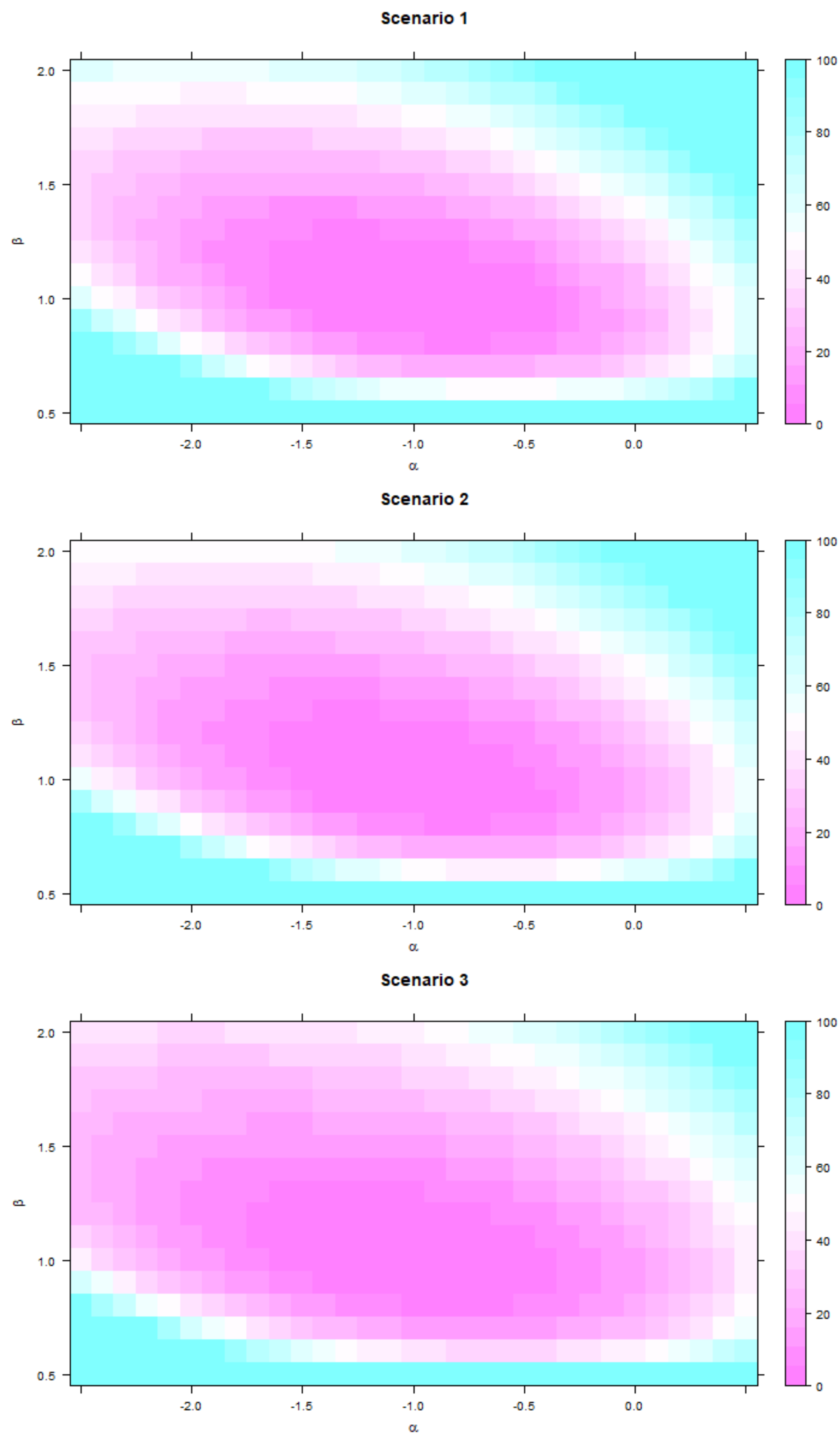

Figure A.13: Gain in efficiency from one-stage to three-stage design

$N = 600$, $(\alpha_t = -1, \beta_t = 1)$

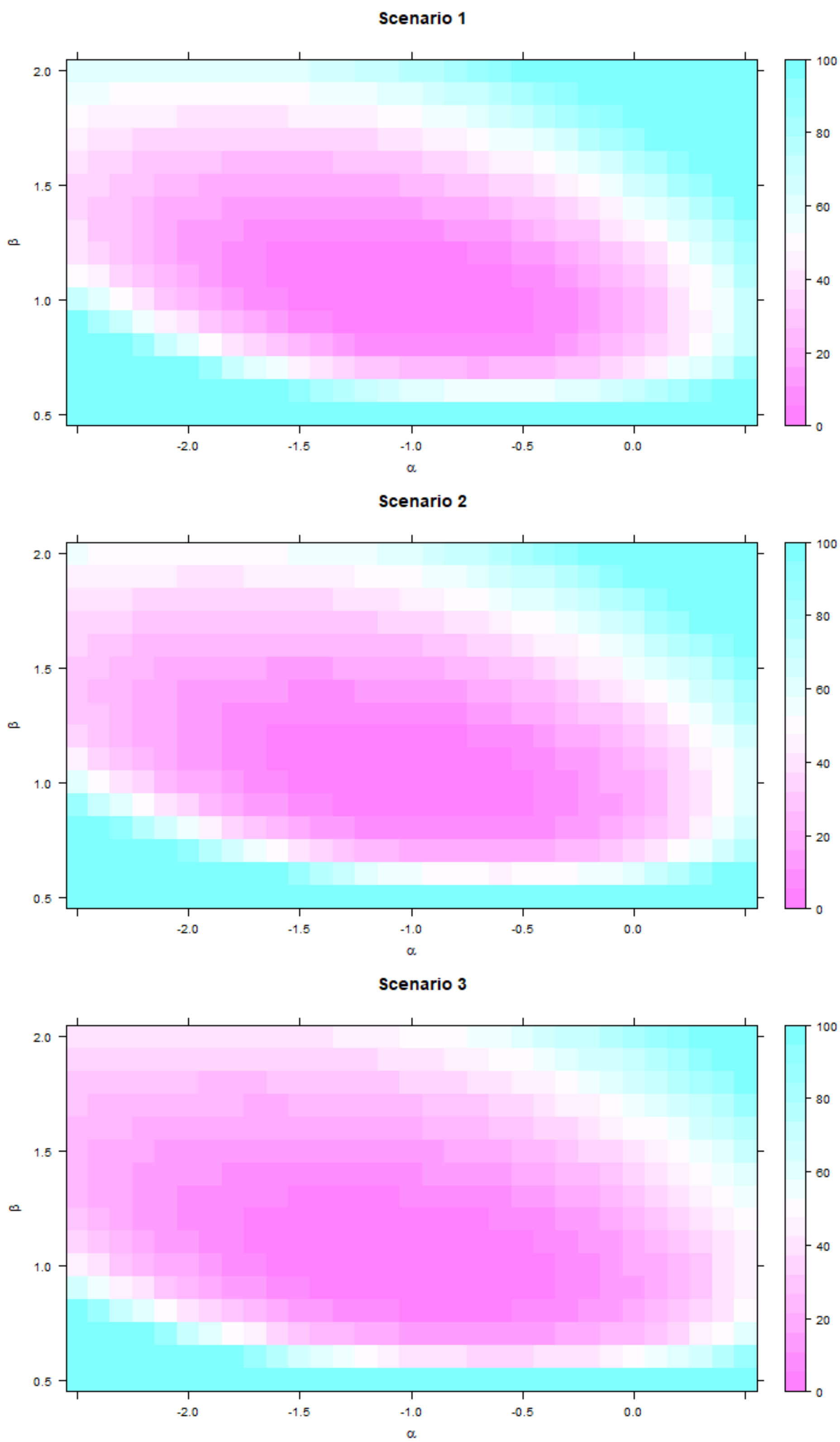

Figure A.14: Gain in efficiency from one-stage to three-stage design

$N = 900$, $(\alpha_t = -1, \beta_t = 1)$

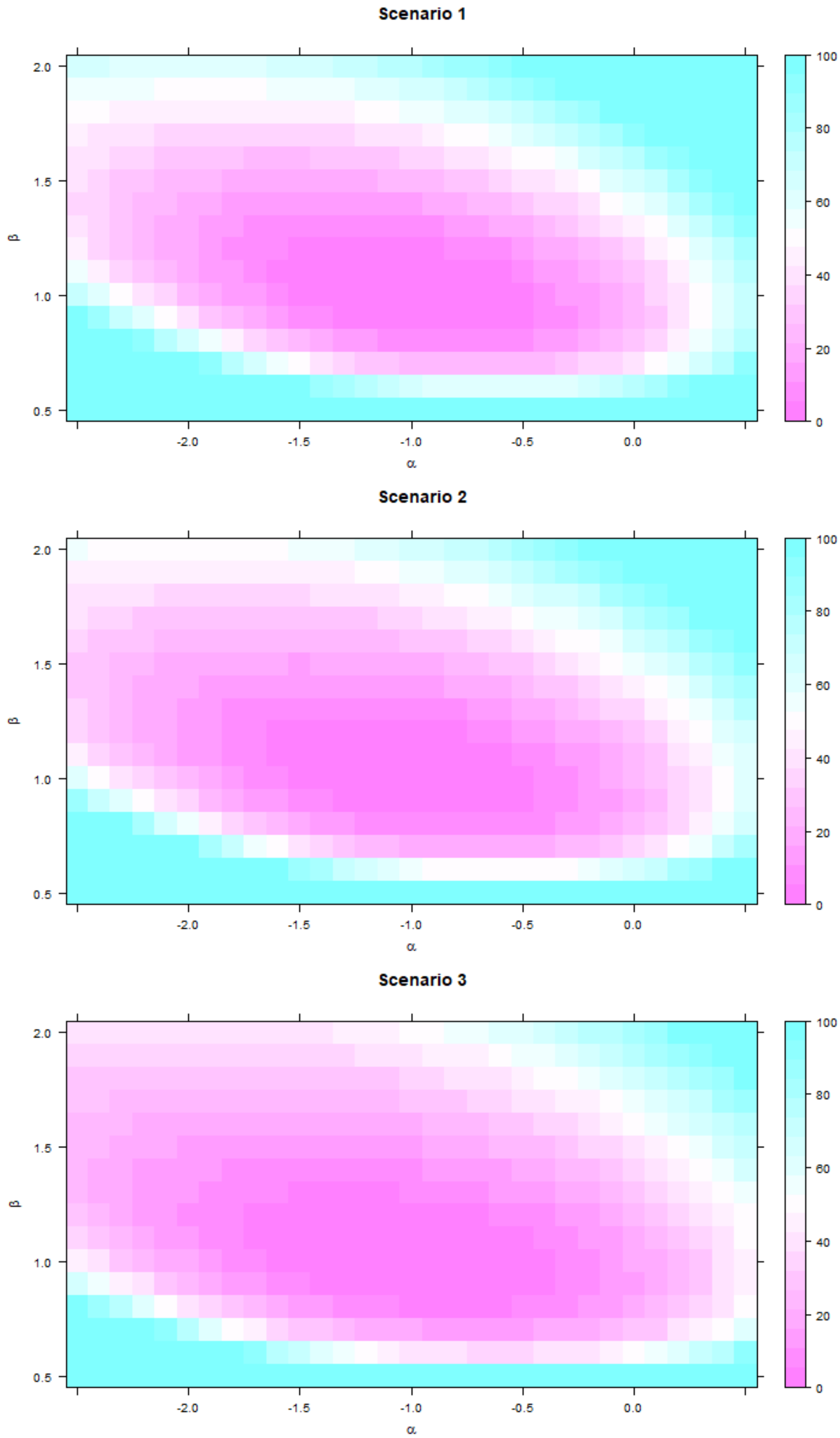

Figure A.15: Gain in efficiency from one-stage to three-stage design

$N = 1200$, $(\alpha_t = -1, \beta_t = 1)$

		Gain in Efficiency (%)		
		Min	Median	Max
N=150	Scenario 1	10.18	20.3	195.4
	Scenario 2	2.1	6.2	130.3
	Scenario 3	0.5	2.0	72.9
N=300	Scenario 1	4.7	10.9	190.3
	Scenario 2	1.1	3.2	125.8
	Scenario 3	0.3	1.0	76.2
N=600	Scenario 1	2.3	5.4	164.7
	Scenario 2	0.6	1.5	121.0
	Scenario 3	0.1	0.4	68.9
N=900	Scenario 1	1.5	3.5	162.5
	Scenario 2	0.3	0.9	108.3
	Scenario 3	0.1	0.3	57.4
N=1200	Scenario 1	1.1	2.5	154.7
	Scenario 2	0.3	0.7	96.2
	Scenario 3	0.1	0.2	48.5

Table A.4: Gain in efficiency from two-stage to three-stage, $(\alpha_t = -1, \beta_t = 1)$

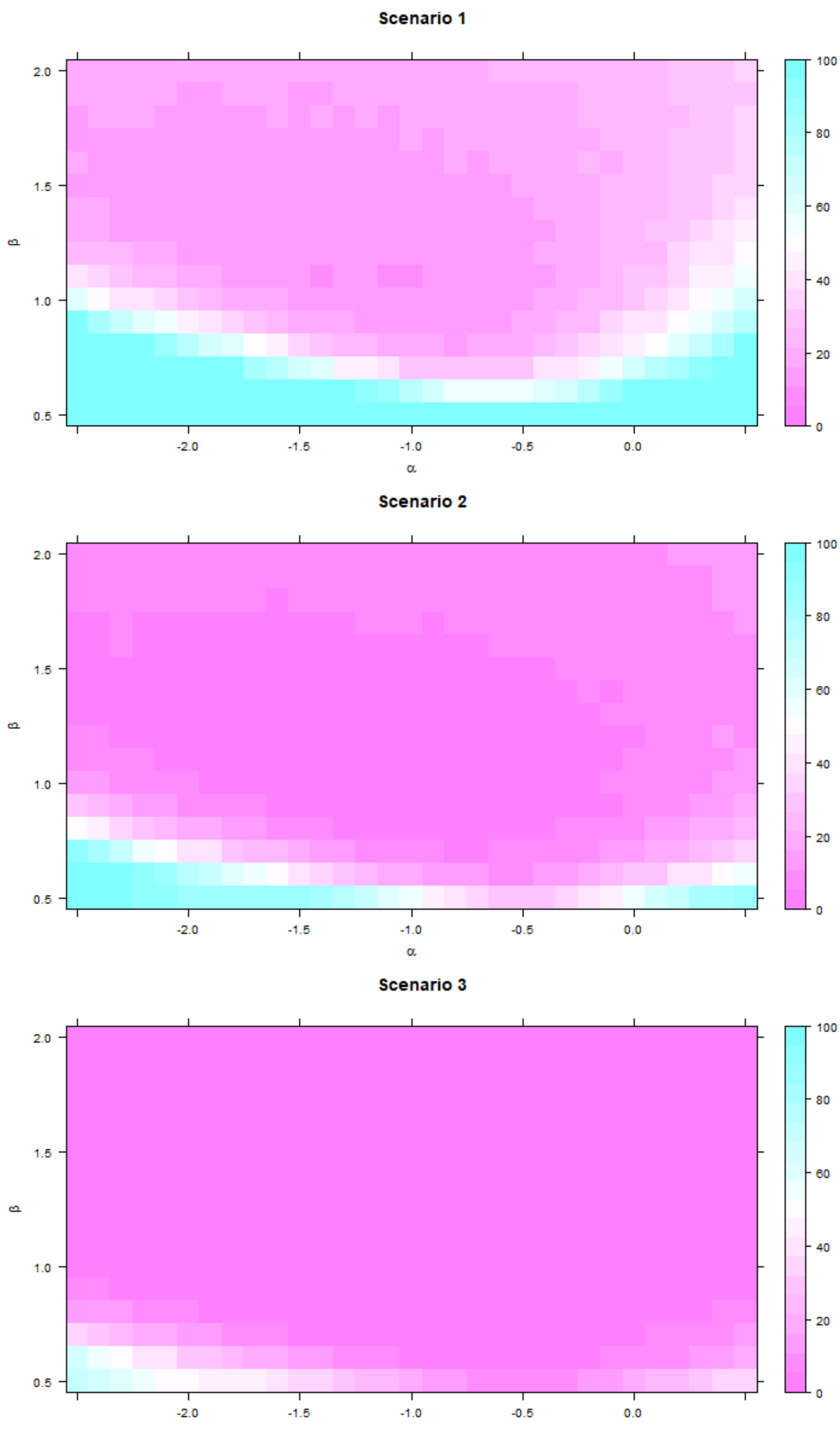

Figure A.16: Gain in efficiency from one-stage to three-stage design

$N = 150$, $(\alpha_t = -1, \beta_t = 1)$

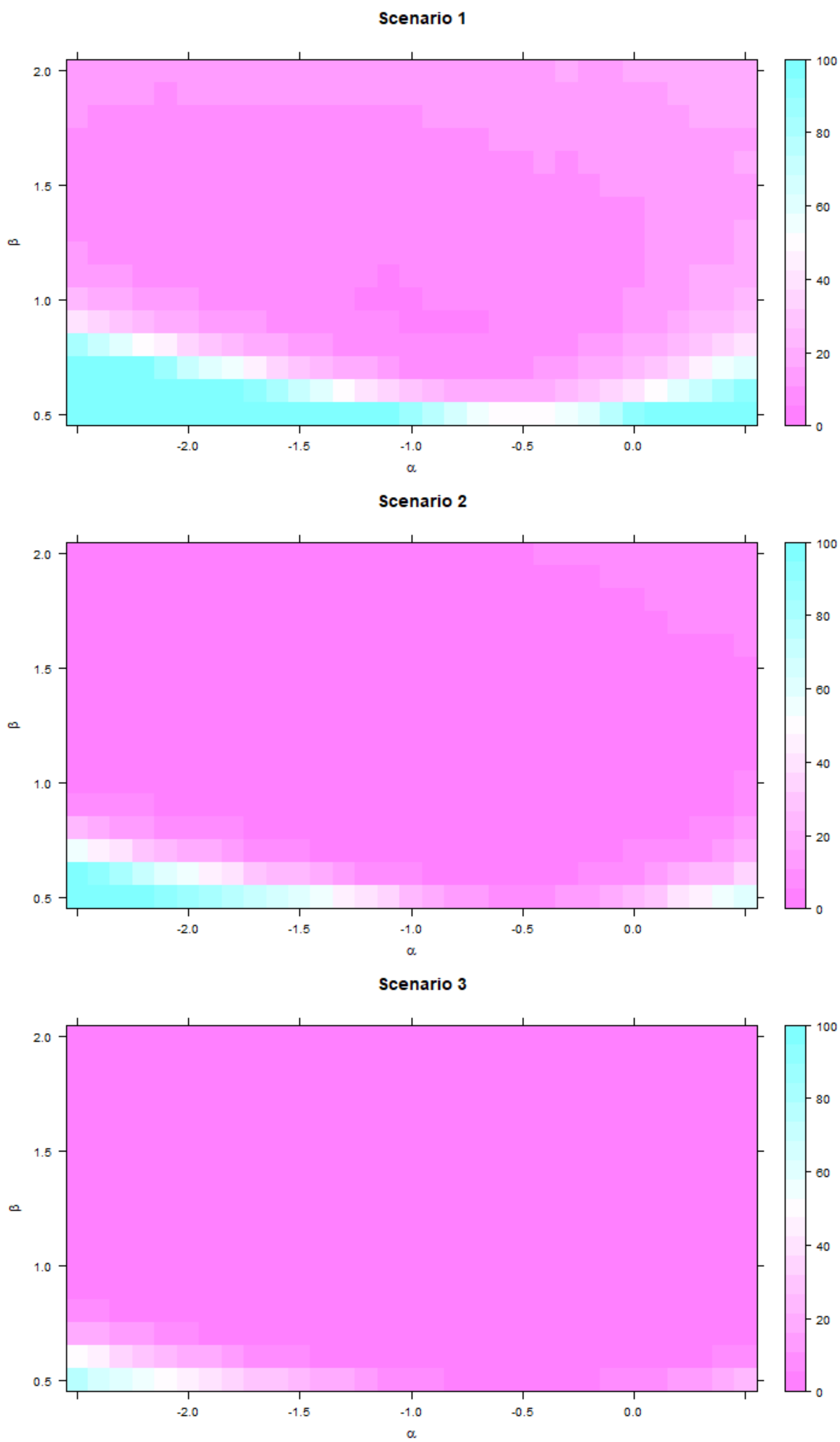

Figure A.17: Gain in efficiency from one-stage to three-stage design

$N = 300$, $(\alpha_t = -1, \beta_t = 1)$

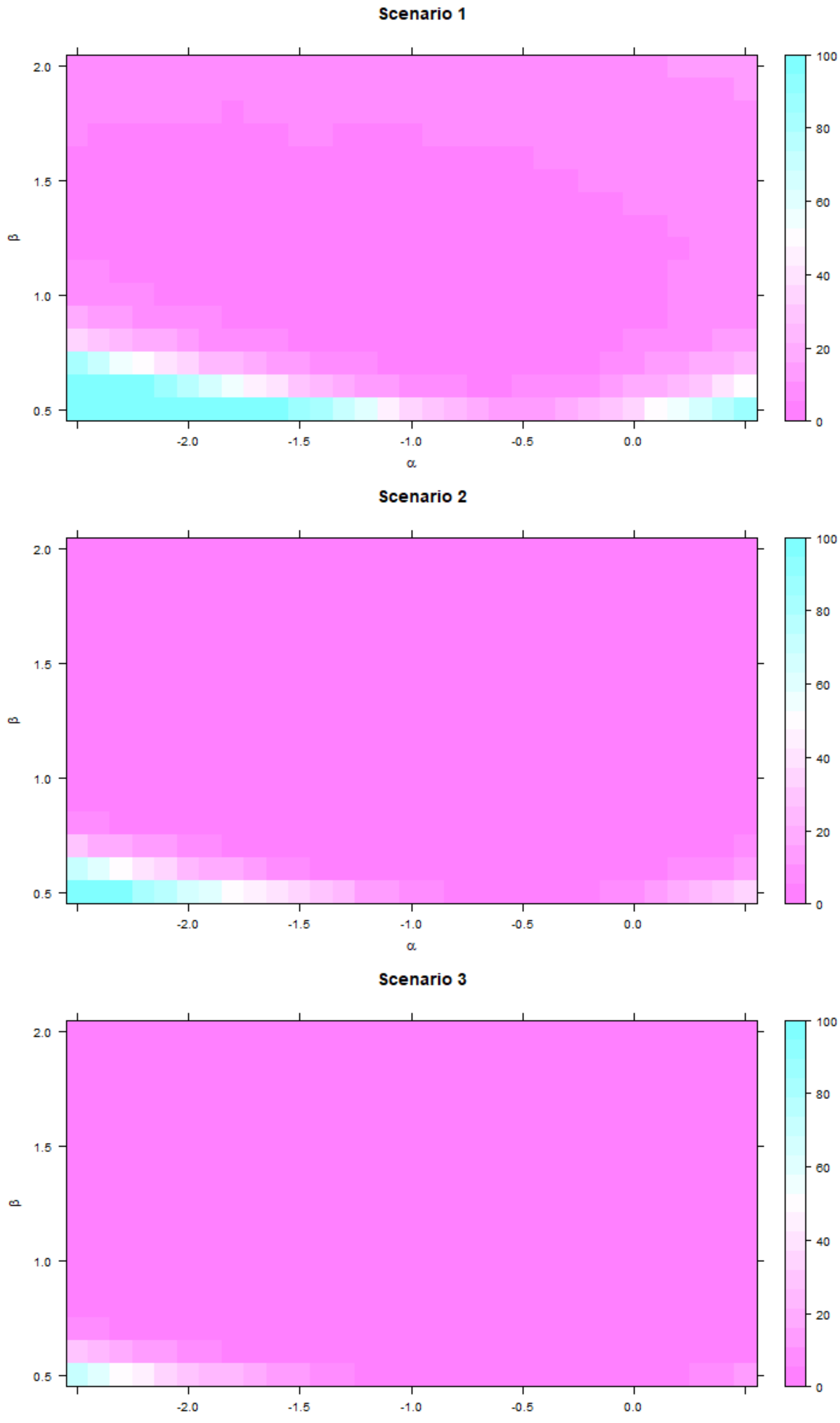

Figure A.18: Gain in efficiency from one-stage to three-stage design

$N = 600$, $(\alpha_t = -1, \beta_t = 1)$

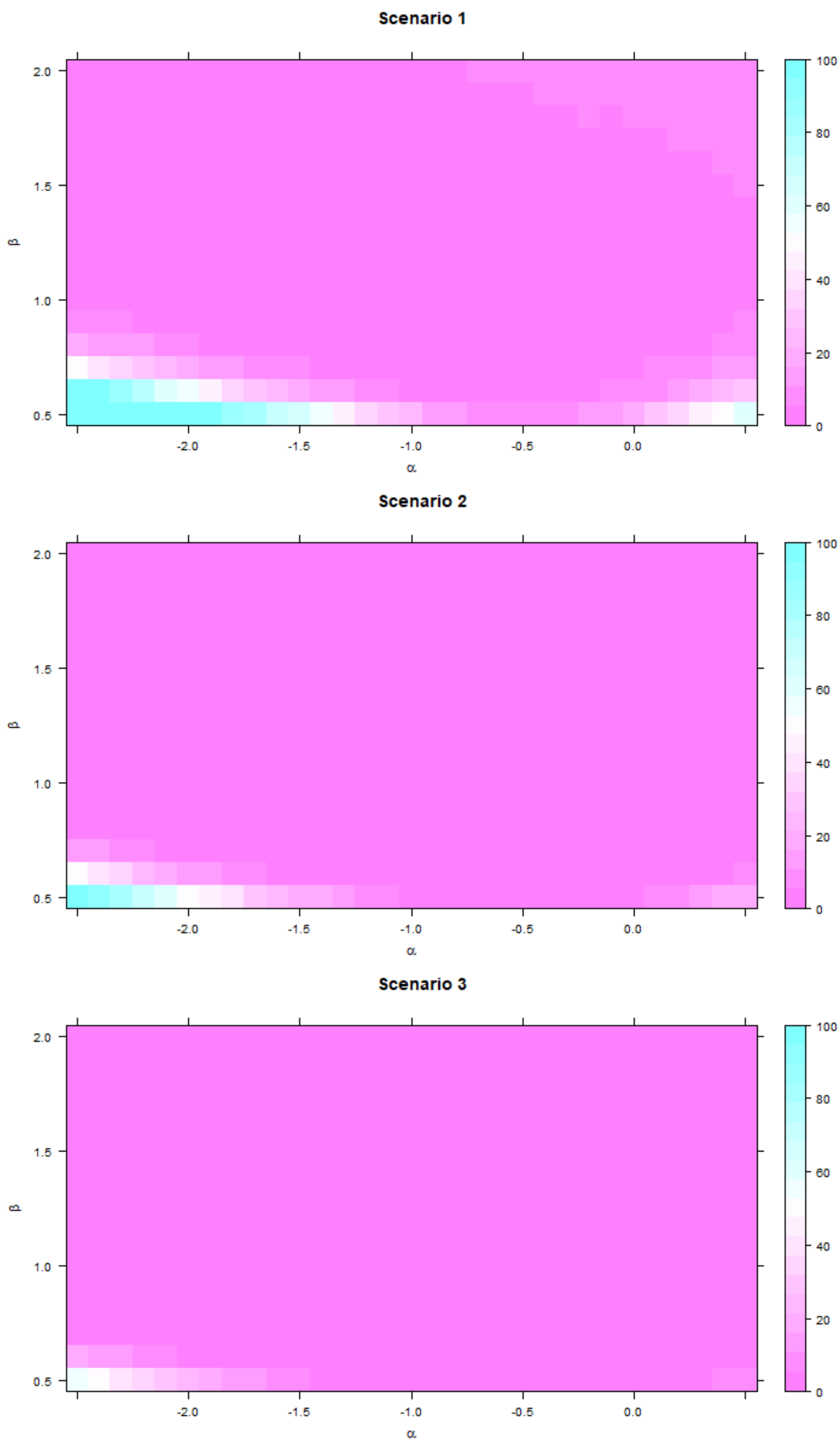

Figure A.19: Gain in efficiency from one-stage to three-stage design

$N = 900$, $(\alpha_t = -1, \beta_t = 1)$

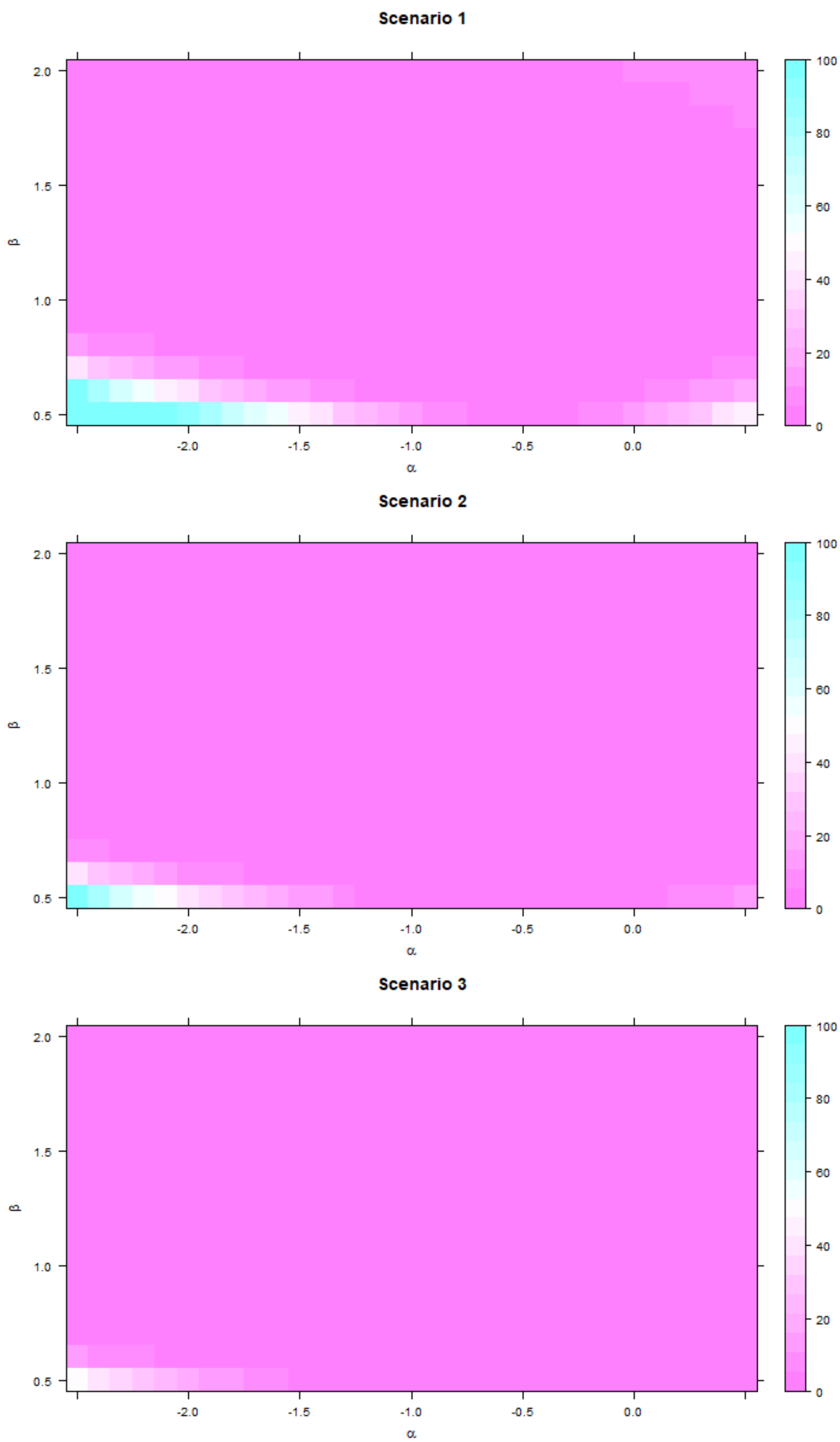

Figure A.20: Gain in efficiency from one-stage to three-stage design

$N = 1200$, $(\alpha_t = -1, \beta_t = 1)$

		Gain in Efficiency (%)		
		Min	Median	Max
N=150	Scenario 1	-20.8	11.6	5768.5
	Scenario 2	-7.2	21.6	4437.9
	Scenario 3	-4.0	21.1	2615.4
N=300	Scenario 1	-7.5	33.0	6659.5
	Scenario 2	-3.5	36.7	4858.6
	Scenario 3	-2.0	29.0	2817.4
N=600	Scenario 1	-3.4	48.2	7289.4
	Scenario 2	-1.7	42.3	5209.5
	Scenario 3	-1.0	32.5	3001.6
N=900	Scenario 1	-2.3	52.3	7536.2
	Scenario 2	-1.2	43.9	5374.2
	Scenario 3	-0.7	33.4	3126.7
N=1200	Scenario 1	-1.8	54.1	7705.0
	Scenario 2	-0.9	44.7	5492.5
	Scenario 3	-0.5	33.8	3225.8

Table A.5: Gain in efficiency from one-stage to three-stage, ($\alpha_t = -2, \beta_t = 1$)

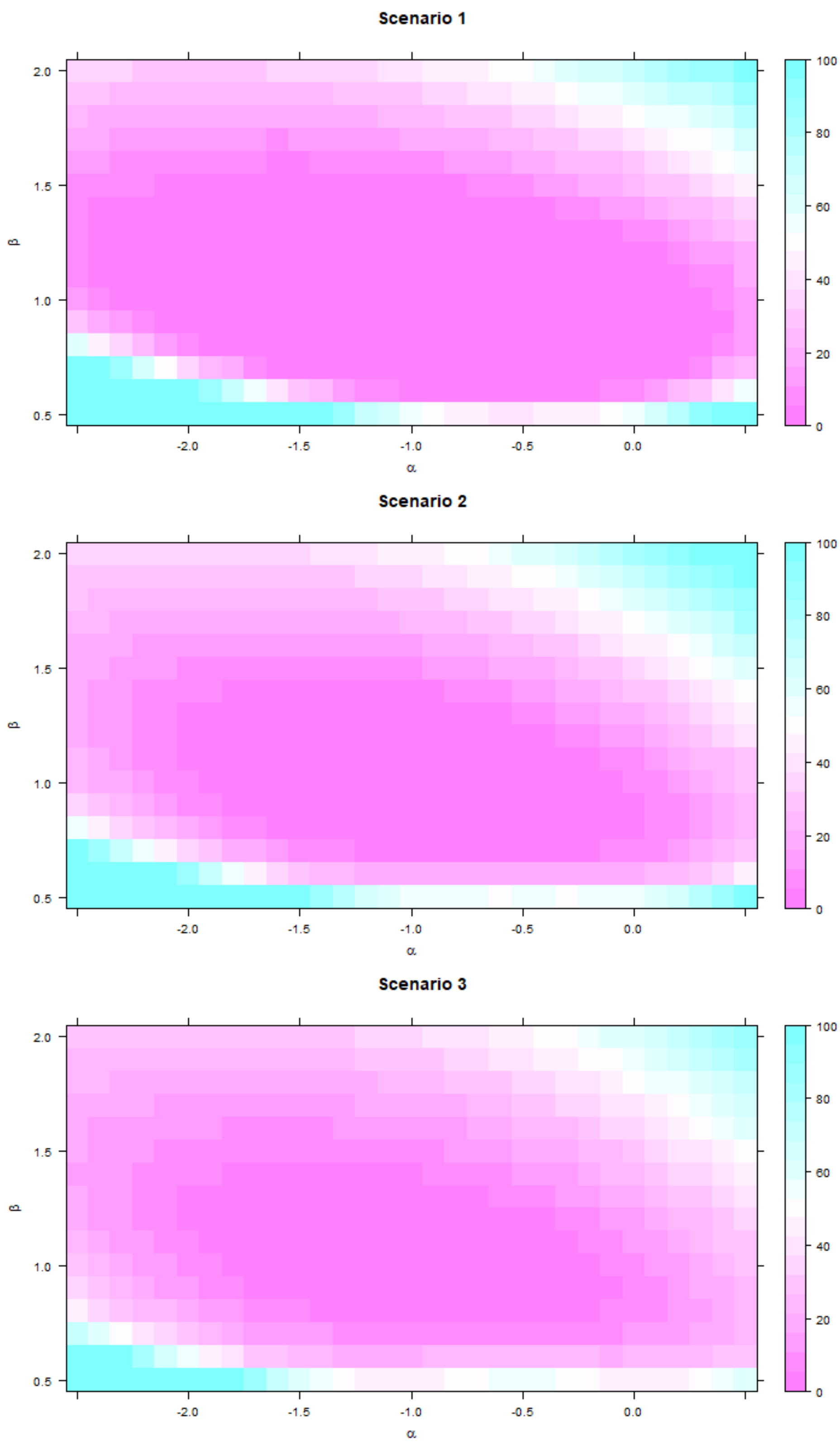

Figure A.21: Gain in efficiency from one-stage to three-stage design

$N = 150$, $(\alpha_t = -2, \beta_t = 1)$

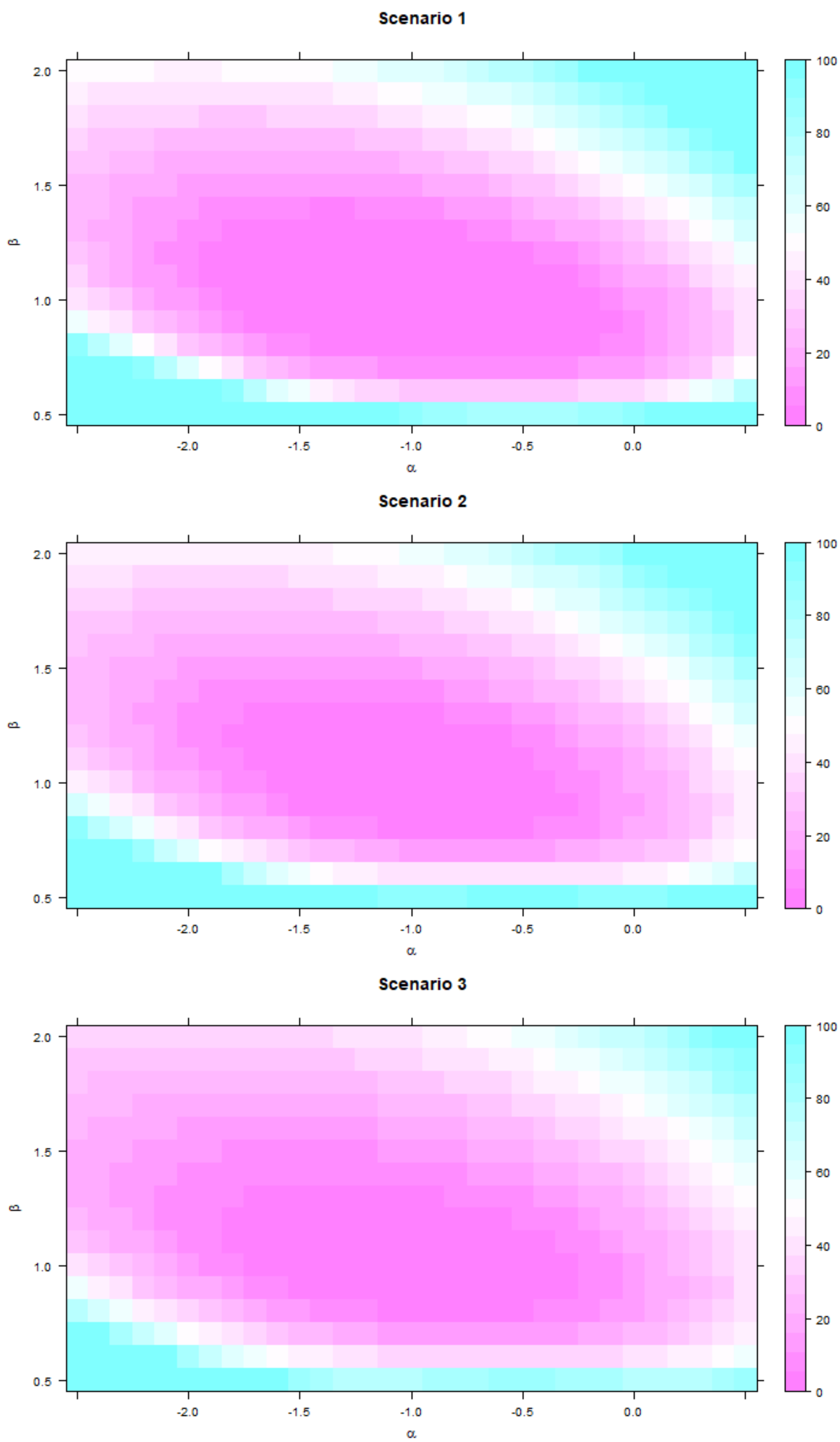

Figure A.22: Gain in efficiency from one-stage to three-stage design

$N = 300$, $(\alpha_t = -2, \beta_t = 1)$

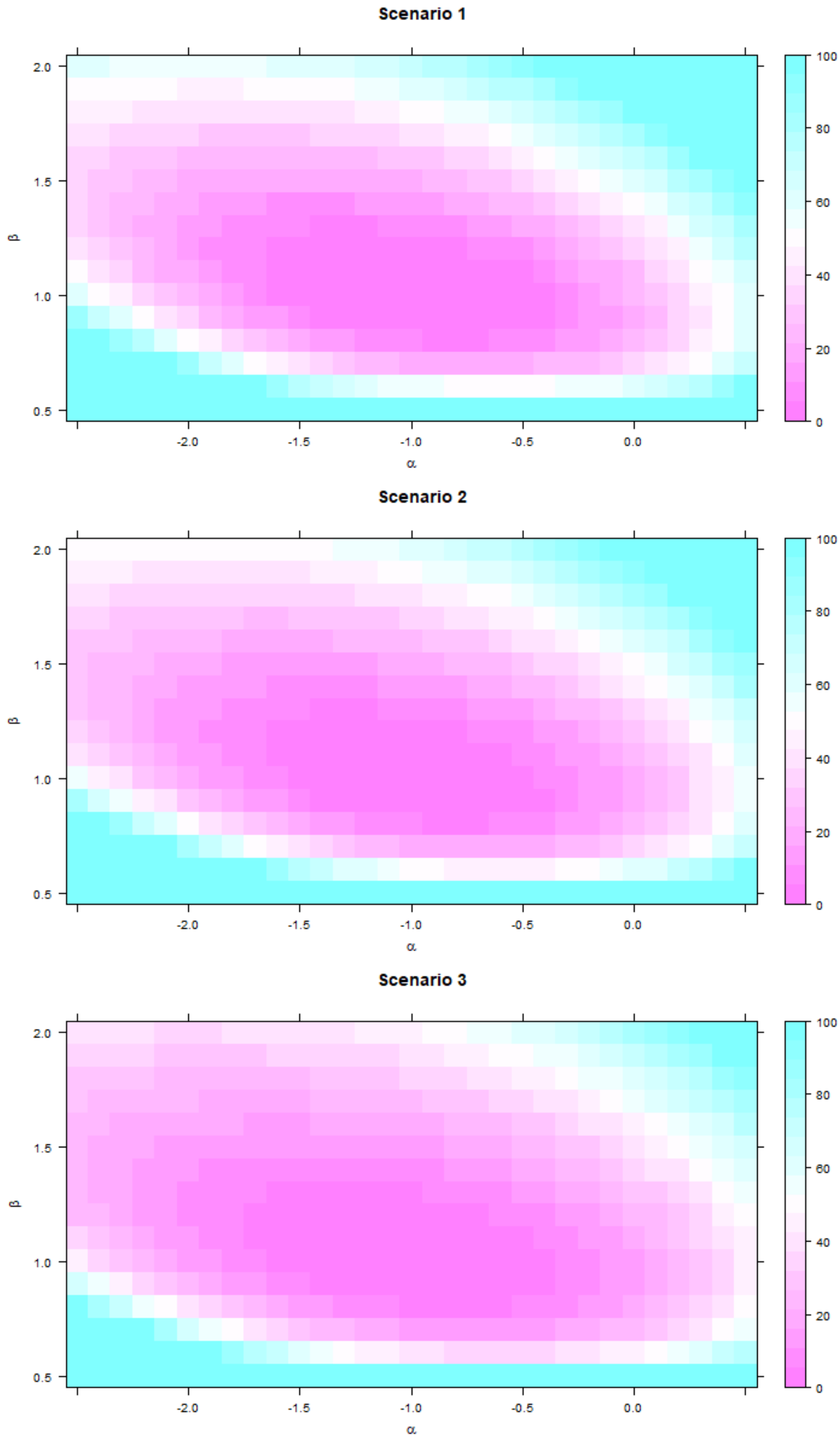

Figure A.23: Gain in efficiency from one-stage to three-stage design

$N = 600$, $(\alpha_t = -2, \beta_t = 1)$

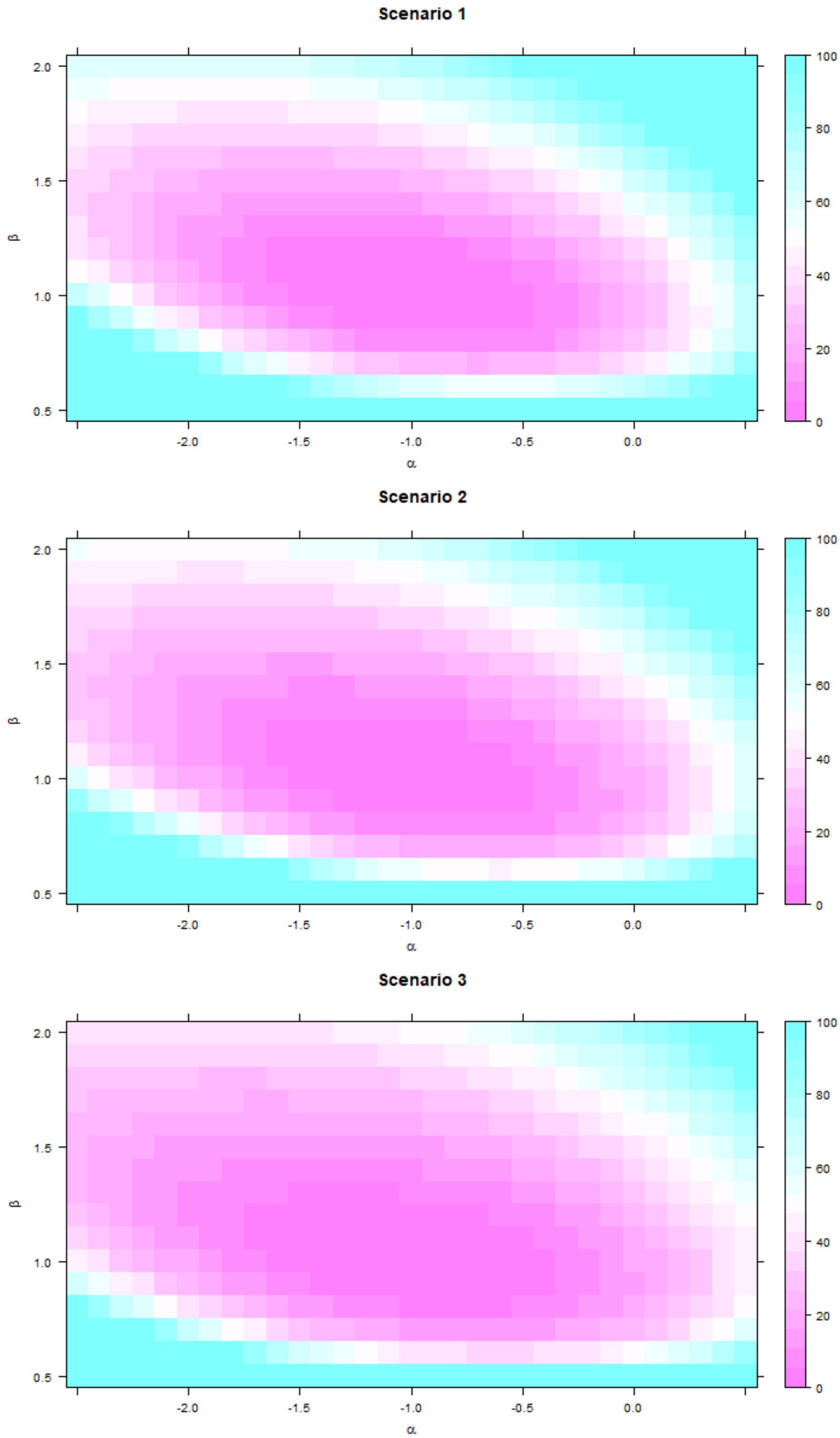

Figure A.24: Gain in efficiency from one-stage to three-stage design

$N = 900$, $(\alpha_t = -2, \beta_t = 1)$

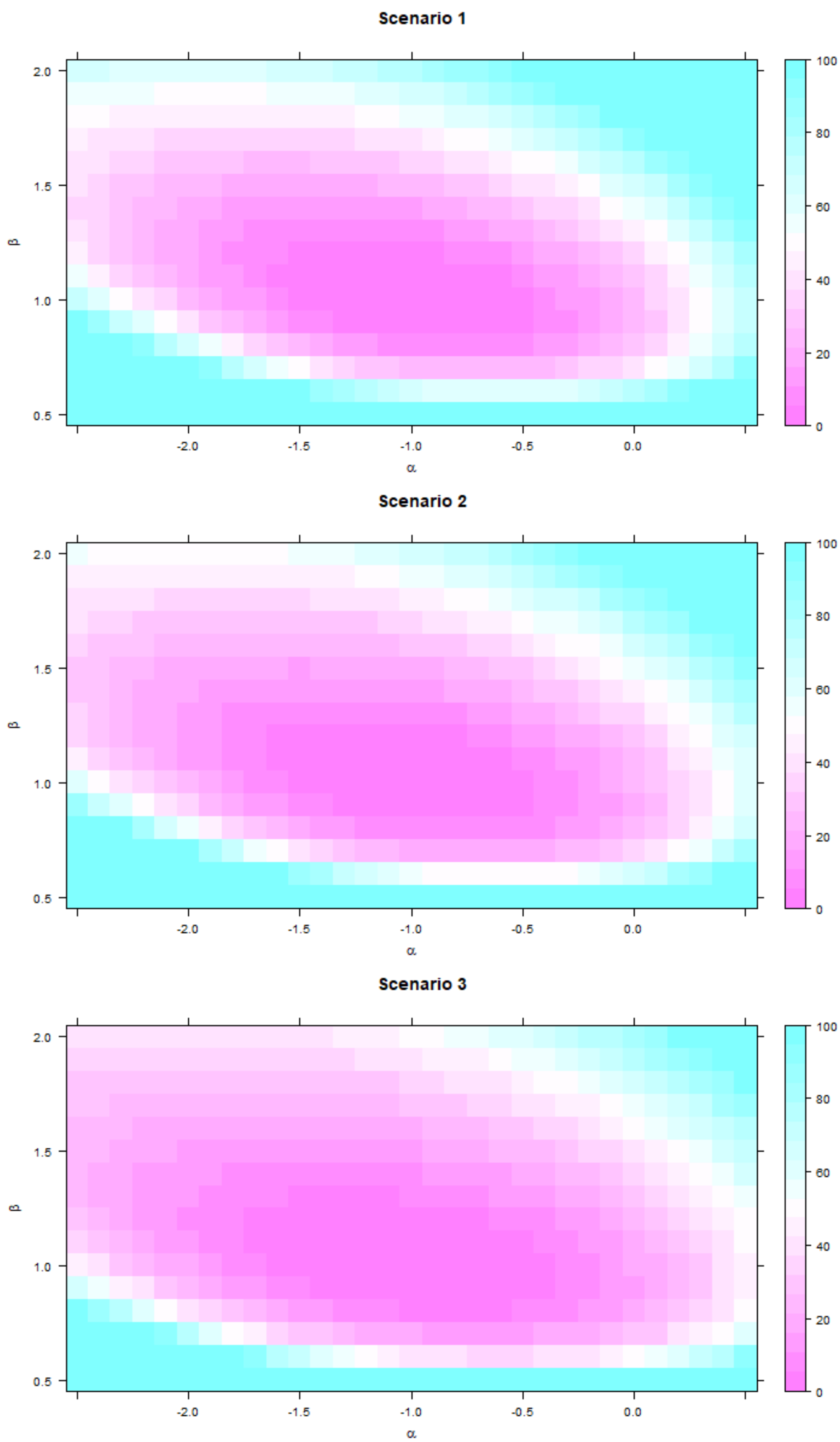

Figure A.25: Gain in efficiency from one-stage to three-stage design

$N = 1200$, $(\alpha_t = -2, \beta_t = 1)$

		Gain in Efficiency (%)		
		Min	Median	Max
N=150	Scenario 1	10.2	20.3	195.4
	Scenario 2	2.1	6.2	130.3
	Scenario 3	0.5	2.0	72.9
N=300	Scenario 1	4.7	10.9	190.3
	Scenario 2	1.1	3.2	125.8
	Scenario 3	0.3	1.0	76.2
N=600	Scenario 1	2.3	5.4	164.7
	Scenario 2	0.6	1.5	121.0
	Scenario 3	0.1	0.4	69.0
N=900	Scenario 1	1.5	3.5	162.5
	Scenario 2	0.4	1.0	108.3
	Scenario 3	0.1	0.3	57.4
N=1200	Scenario 1	1.1	2.5	154.8
	Scenario 2	0.27	0.7	96.2
	Scenario 3	0.1	0.2	48.5

Table A.6: Gain in efficiency from two-stage to three-stage, ($\alpha_t = -2, \beta_t = 1$)

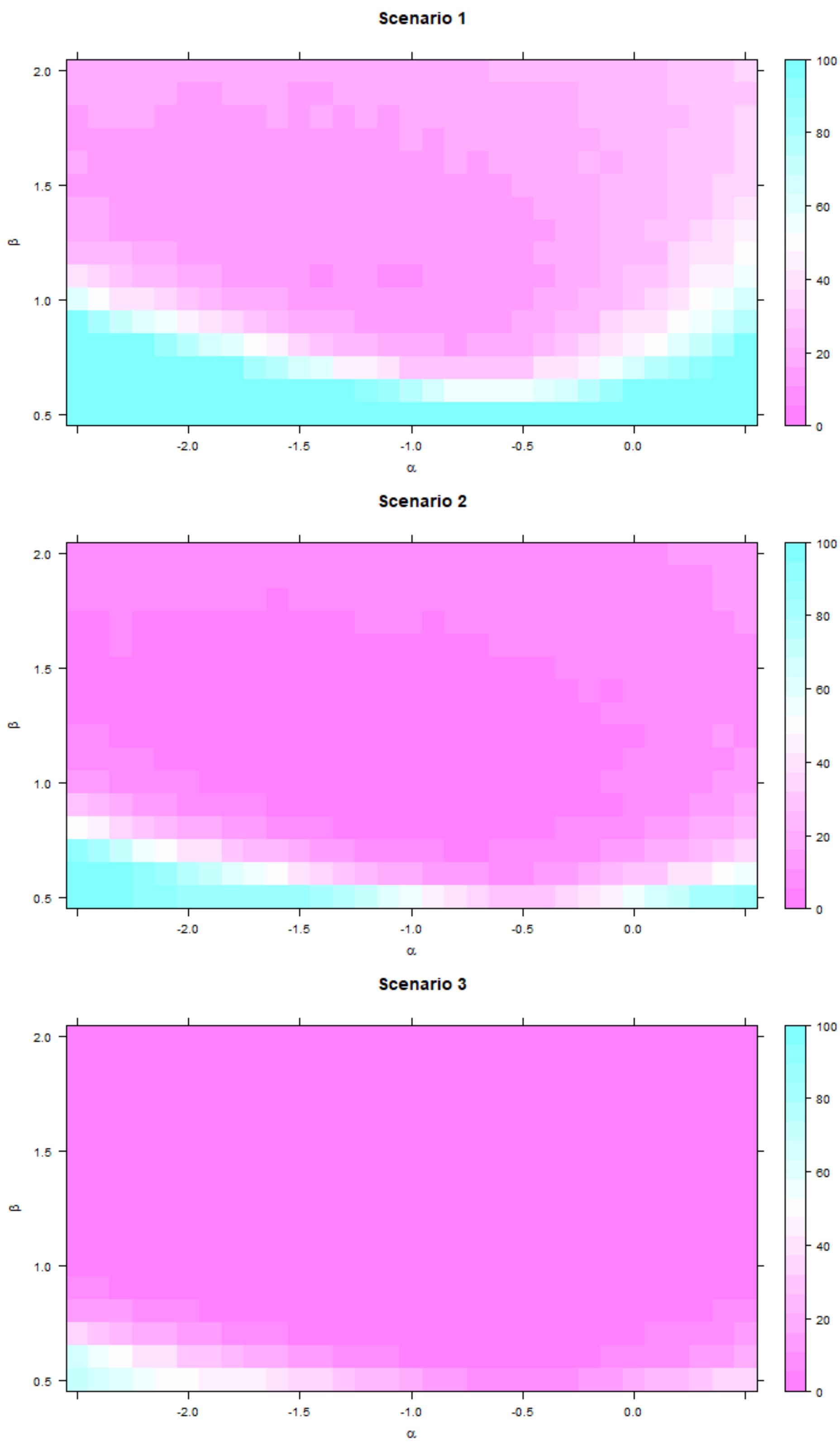

Figure A.26: Gain in efficiency from one-stage to three-stage design

$N = 150$, $(\alpha_t = -2, \beta_t = 1)$

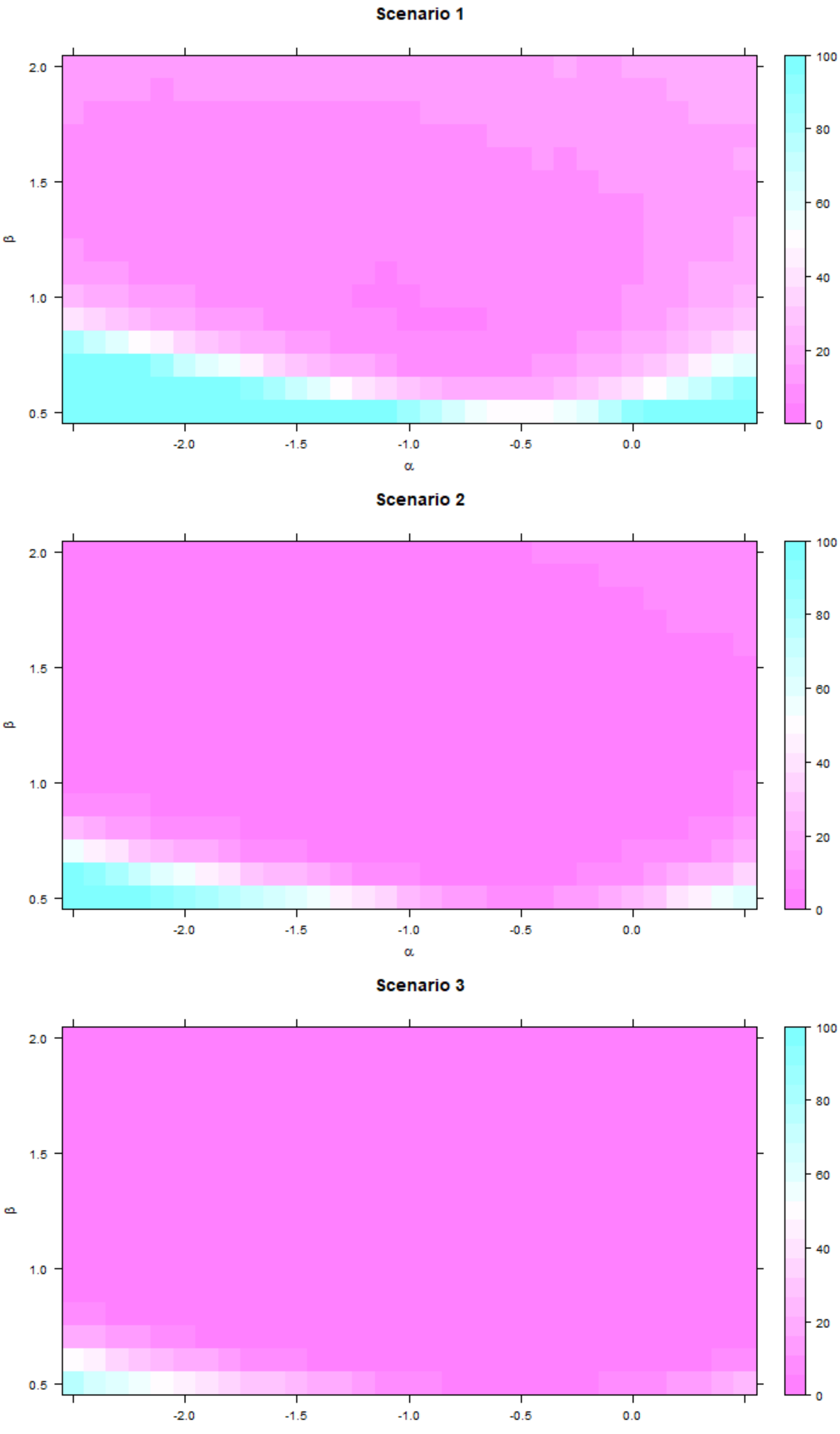

Figure A.27: Gain in efficiency from one-stage to three-stage design

$N = 300$, $(\alpha_t = -2, \beta_t = 1)$

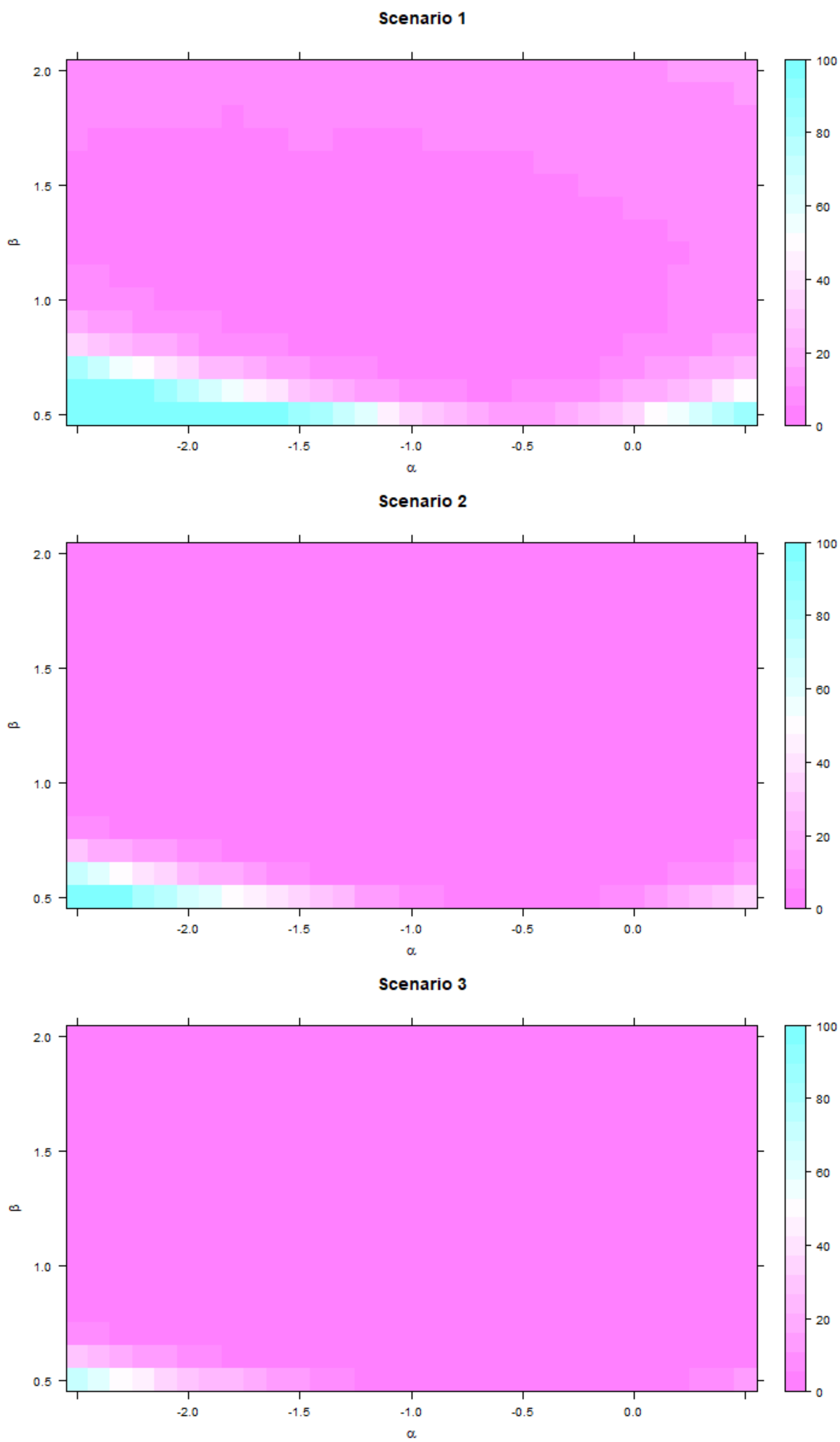

Figure A.28: Gain in efficiency from one-stage to three-stage design

$N = 600$, $(\alpha_t = -2, \beta_t = 1)$

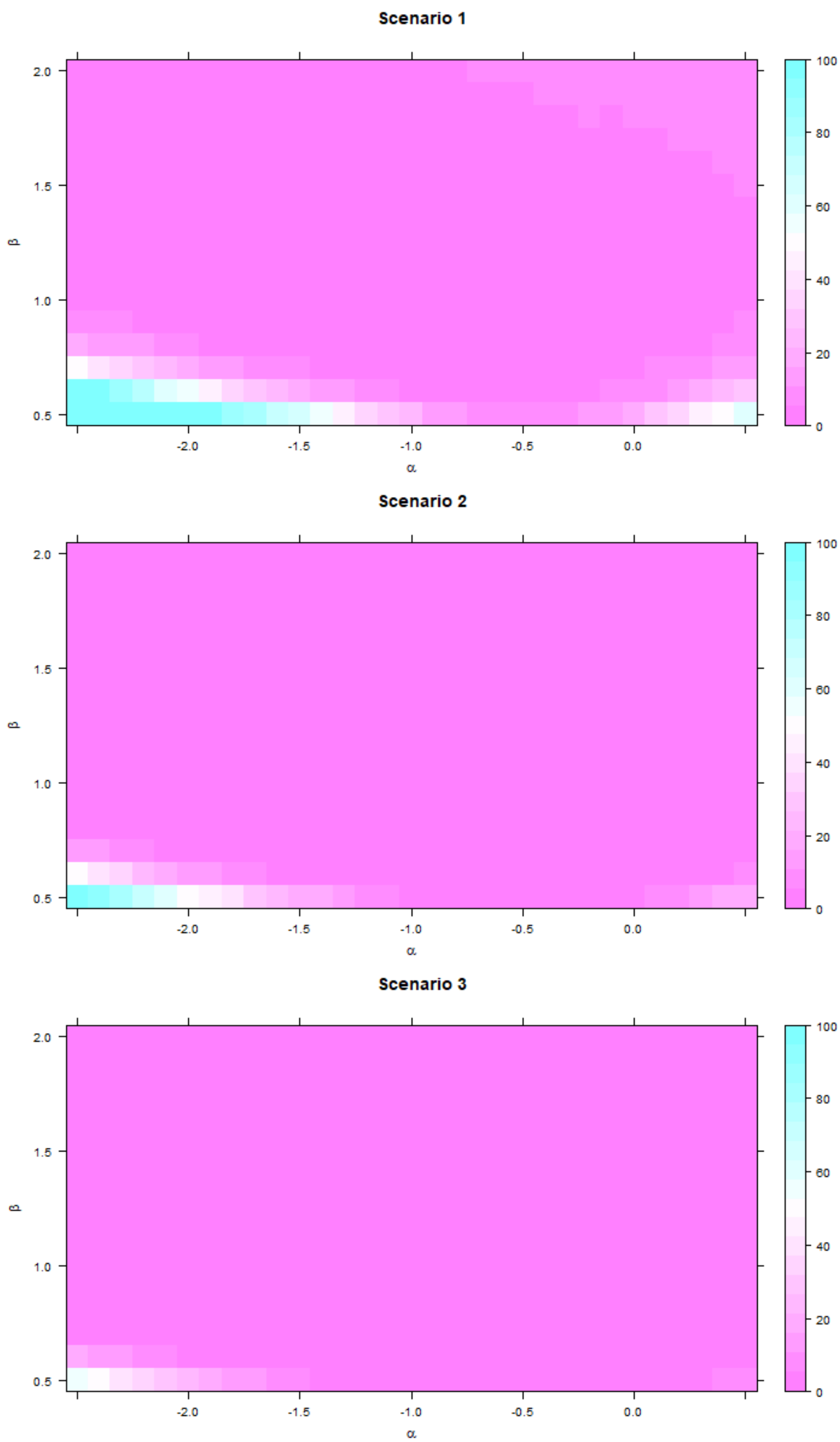

Figure A.29: Gain in efficiency from one-stage to three-stage design

$N = 900$, $(\alpha_t = -2, \beta_t = 1)$

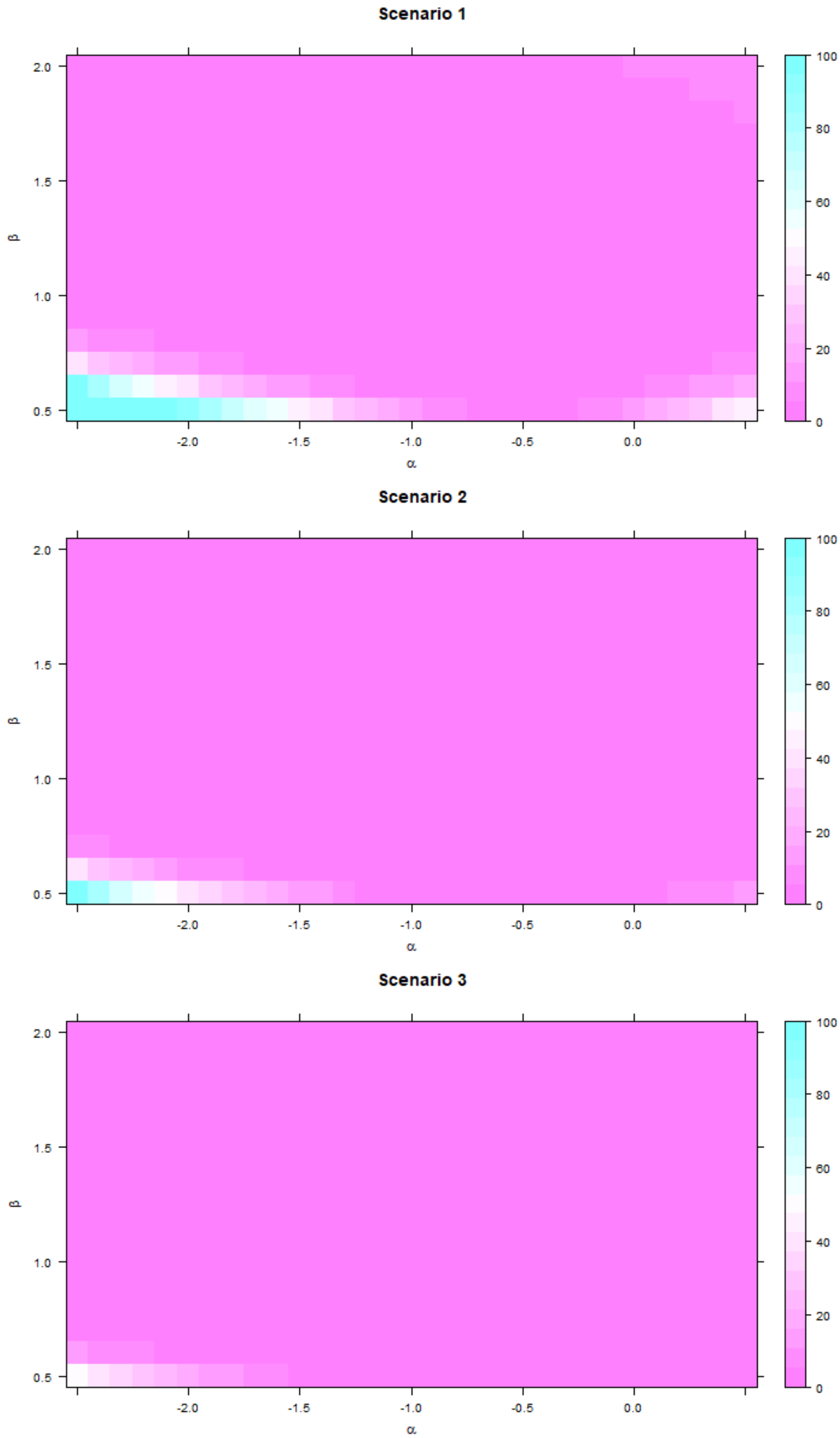

Figure A.30: Gain in efficiency from one-stage to three-stage design

$N = 1200$, $(\alpha_t = -2, \beta_t = 1)$

Bibliography

Box, G. E., & Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(1), 1–38.

Cope, A. P. (2019). Considerations for optimal trial design for rheumatoid arthritis prevention studies. *Clinical therapeutics*.

Fedorov, V. V., & Leonov, S. L. (2013). *Optimal design for nonlinear response models*. CRC Press.

Fisher, R. A. (1949). *The design of experiments*. Oliver & Boyd.

Fredrickson, D. D., Jones, T. L., Molgaard, C. A., Carman, C. G., Schukman, J., Dismuke, S. E., & Ablah, E. (2005). Optimal design features for surveying low-income populations. *Journal of health care for the poor and underserved*, *16*(4), 677–690.

Heise, M. A., & Myers, R. H. (1996). Optimal designs for bivariate logistic regression. *Biometrics*, 613–624.

Khan, M., & Yazdi, A. (1988). On d-optimal designs for binary data. *Journal of statistical planning and inference*, *18*(1), 83–91.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, *21*(2), 272–304.

Kiefer, J. (1961). Optimum designs in regression problems, ii. *The Annals of Mathematical Statistics*, *32*(1), 298–325.

Kiefer, J., & Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, *12*, 363–366.

Klick, B., Nishiura, H., & Cowling, B. J. (2012). Optimal design of intervention studies to prevent influenza in healthy cohorts. *PloS one*, *7*(4), e35166.

Klick, B., Nishiura, H., Leung, G., & Cowling, B. (2014). Optimal design of studies of influenza transmission in households. ii: comparison between cohort and case-ascertained studies. *Epidemiology & Infection*, *142*(4), 744–752.

Mathew, T., & Sinha, B. K. (2001). Optimal designs for binary data under logistic regression. *Journal of Statistical Planning and Inference*, *93*(1-2), 295–307.

Minkin, S. (1987). Optimal designs for binary data. *Journal of the American Statistical Association*, *82*(400), 1098–1103.

Nandy, K., & Nandy, R. R. (2015). A study of the performance of 2-stage adaptive optimal designs in a logistic dose-response model. *Communications in Statistics-Simulation and Computation*.

Pukelsheim, F. (1993). Optimal design of experiments.© john wiley and sons. *Inc., New York, New York*.

Santos, A. R., da Silva, A. F., Batista, A. F., Freitas, C. F., Bona, E., Sereia, M. J., ... Mikcha, J. M. (2020). Application of response surface methodology to evaluate photodynamic inactivation mediated by eosin y and 530 nm led against staphylococcus aureus. *Antibiotics*, *9*(3), 125.

Shah, K. R., & Sinha, B. K. (1989). Theory of optimal designs. Springer.

Silvey, S. (1980). *Optimal design: an introduction to the theory for parameter estimation* (Vol. 1). Springer Science & Business Media.

Sitter, R., & Wu, C. (1993). Optimal designs for binary response experiments: Fieller, d, and a criteria. *Scandinavian Journal of Statistics*, 329–341.

White, L. V. (1973). An extension of the general equivalence theorem to nonlinear models. *Biometrika*, *60*(2), 345–348.

Wu, C. J. (1985). Efficient sequential designs with binary data. *Journal of the American Statistical Association*, *80*(392), 974–984.

Yang, M. (2008). A-optimal designs for generalized linear models with two parameters. *Journal of Statistical Planning and Inference*, *138*(3), 624–641.