

**Credible Subgroups: Identifying the Population that
Benefits from Treatment**

**A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY**

Patrick Martin Schnell

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy**

Advised by Bradley P. Carlin, Ph.D

May, 2017

© Patrick Martin Schnell 2017
ALL RIGHTS RESERVED

Acknowledgements

I owe so much to my parents, Dan and Beverly, for providing me with every opportunity to develop and pursue my passions, and to my wife, Mary Kate, for taking the leap of faith to move to the Twin Cities with me and share this great journey. My advisor, Brad Carlin, provided endless advice and support in the unfamiliar world of academic research, and Jim Hodges never let me lose sight of my ideals; together their high expectations and ample guidance helped me to grow as a scholar. Sudipto Banerjee, Joe Koopmeiners, and Ann Brearley also served, sometimes from a distance, as role models and exemplars of facets of my career-to-come. Qi Tang and Peter Müller contributed many helpful insights into the needs of treatment developers and ideas in decision theory, respectively, and helped in writing the papers that make up the bulk of this dissertation.

Several sources provided financial support throughout the work contained in this dissertation: the Division of Biostatistics Research Training Fellowship, the National Cancer Institute, AbbVie, and the University of Minnesota Doctoral Dissertation Fellowship.

Dedication

To those overlooked in medical research.

Abstract

A single treatment may have a different effect on different patients. In particular, some patients may benefit from a given treatment while others do not. Often, some of the variation in effect among patients can often be explained by characteristics of those patients that are observable before treatment. Widespread acknowledgment of treatment effect variation due to observable patient characteristics has increased the health science community's interest in a broad field referred to as personalized or precision medicine. Among the aims of precision medicine are identifying the set of treatments that would benefit a given patient, and conversely, identifying the population of patients who would benefit from a given treatment. We treat the latter problem in the context of clinical trials run by treatment developers (e.g., pharmaceutical companies), with special attention paid to interactions between those developers and the relevant regulatory agencies (e.g., the US Food and Drug Administration). The primary difficulty in estimating the benefiting population in such settings is controlling the frequency with which at least one type of patient is incorrectly determined to benefit, and doing so in a way that does not render the approach excessively conservative.

As a motivating application throughout this dissertation, we consider a battery of related clinical trials of treatments for Alzheimer's disease carried out by the pharmaceutical company AbbVie. These trials contain a small number of continuous and binary baseline patient characteristics that may influence the treatment effect. We apply standard and more novel regression models to the supplied data and develop methods of inference to accommodate the varied features of the datasets, such as nonlinear effects, multiple important endpoints, more than two treatments, and regions of the covariate space that are sparse in observations or lacking common support among treatment arms. We also discuss topics in practical implementation of these methods. Our approaches yield reliable and easily interpretable inferences regarding the population that benefits from treatment.

Contents

Acknowledgements	i
Dedication	ii
Abstract	iii
List of Tables	viii
List of Figures	x
1 Identifying the Benefiting Subgroup	1
1.1 Addressing treatment effect heterogeneity	1
1.2 Modes of inference for personalized medicine	3
1.2.1 One patient, many interventions	4
1.2.2 Many patients, one intervention	5
1.2.3 Many patients, many interventions	8
1.2.4 Frequentist or Bayesian?	8
1.3 Clinical motivation and plan of dissertation	9
1.3.1 Alzheimer’s disease	9
1.3.2 Simple add-on therapy dataset	10
1.3.3 Multi-endpoint dataset	10
1.3.4 Multi-trial dataset	10
1.3.5 Implementation and future work	11

2	Credible Subgroups	12
2.1	Bounds as estimators	12
2.1.1	A normal hierarchical linear model	15
2.1.2	Highest posterior density credible subgroups	16
2.1.3	Restricted covariate spaces	17
2.1.4	Purely Bayesian credible subgroups	17
2.1.5	Asymptotic properties of credible subgroups	18
2.1.6	A step-down multiple testing procedure	19
2.1.7	Maximum credible levels	20
2.2	Simulations	22
2.3	Analysis of simple add-on therapy dataset	26
2.4	Discussion	30
3	Credible Subgroups for General Regression Models	32
3.1	General forms of simultaneous credible bands	33
3.2	Parametric regression	34
3.2.1	Generalized linear models	34
3.2.2	Variable selection	36
3.3	Semiparametric and nonparametric regression	38
3.3.1	Simulations	40
3.3.2	Semiparametric example analysis	42
3.4	Discussion	47
4	Subgroup Inference with Multiple Endpoints and Many Treatments	48
4.1	Subgroup inference	50
4.1.1	Multiple endpoints	50
4.1.2	Multiple endpoints and many treatments	52
4.1.3	Credible subgroups	53
4.2	Simulations	55
4.3	Analysis of multi-endpoint dataset	58
4.4	Discussion	63

5	Considerations for Practical Implementation	65
5.1	Power computations	66
5.1.1	Local power	66
5.1.2	Choice of covariate space	67
5.2	Monte Carlo precision	70
5.3	Diagnostics	74
5.4	Reporting	76
5.5	Software	79
5.5.1	A nonparametric example	81
5.5.2	A parametric example	84
6	Conclusion	89
6.1	Summary of developments	89
6.2	Significance of work	90
6.3	Future work	92
6.3.1	A “ 2α conjecture”	92
6.3.2	A bootstrap counterpart	93
6.3.3	Large numbers of binary covariates	94
6.3.4	Sequential and adaptive trial designs	94
	References	96
	Appendix A. Proofs of theorems	102
A.1	Theorems for simultaneous credible bands	102
A.2	Theorems for sequential testing procedures	105
	Appendix B. Expanded simulation results	107
B.1	Additional simulations for basic credible subgroups	107
B.2	Comparison of parametric, semiparametric, and nonparametric regression	107
	Appendix C. Software Code	117
C.1	credsubs.R	117
C.2	Shiny Calculator App	130
C.2.1	server.R	130

C.2.2	ui.R	132
Appendix D. Acronyms and Symbols		133
D.1	Acronyms	133
D.2	Symbols	134

List of Tables

2.1	Average summary statistics for 80% credible subgroup pairs as well as pointwise (PW) method (n=40). Statistics are averaged without undefined values, e.g. sensitivity of D when B is empty. Coverage rates at or above 80% and low pair sizes (analogous to interval lengths for interval estimation) are desired.	24
2.2	Posterior summaries of selected effect parameters. Continuous covariates are standardized. Estimates greater than 1.96 posterior standard deviations from 0 are marked significant.	28
3.1	Top five models.	37
3.2	Simulation study results. Operating characteristics of 80% credible subgroups with $n = 100$ patients in each study arm. Struck-through sensitivities indicate insufficient coverage and should be treated with caution.	41
5.1	Applicability of behavior arguments to primary functions.	80
B.1	Coverage and model fit statistics for 80% credible subgroup pairs (n=40).	108
B.2	Coverage and model fit statistics for 80% credible subgroup pairs (n=100).	109
B.3	Coverage and model fit statistics for 80% credible subgroup pairs (n=350).	110
B.4	Diagnostic properties of 80% credible subgroup pairs (n=40).	111
B.5	Diagnostic properties of 80% credible subgroup pairs (n=100).	112
B.6	Diagnostic properties of 80% credible subgroup pairs (n=350).	113

B.7	Simulation study results. Operating characteristics of 80% credible subgroups with $n = 25$ patients in each study arm. Struck-through sensitivities indicate insufficient coverage and should be treated with caution.	114
B.8	Simulation study results. Operating characteristics of 80% credible subgroups with $n = 50$ patients in each study arm. Struck-through sensitivities indicate insufficient coverage and should be treated with caution.	115
B.9	Simulation study results. Operating characteristics of 80% credible subgroups with $n = 75$ patients in each study arm. Struck-through sensitivities indicate insufficient coverage and should be treated with caution.	116

List of Figures

2.1	Interpretation of the trichotomy of the covariate space induced by the credible subgroup pair (D, S) relative to the true benefiting subgroup B .	13
2.2	Diagnostic measure comparison in a case with a binary covariate-treatment interaction. Sensitivity (<i>left</i>) and specificity (<i>right</i>) of D in the case $\gamma = (0, 1, 0)$ (treatment effect is determined by a binary covariate). The multiplicity-correcting methods (HPD, RCS, and to a lesser extent PB) maintain extremely high specificity at the expense of sensitivity, especially for small sample sizes. Because the benefit is positive in one group and zero in its complement, the sensitivities of all methods approach 100% for large sample sizes while the specificities remain approximately constant.	25
2.3	Step-down RCS credible subgroups at the 80% level with $\delta = 0$ (left), and at the 50% level with $\delta = 2$ (right). Patients in the control arm are represented by \times and those in the treatment arm by $+$.	29
2.4	Contours of posterior mean and standard error surfaces, which combine to produce credible subgroups.	30
3.1	Observed covariate points (left, $+$ for Donepezil, \times for placebo) and maximum credible level contours for a non-inferiority (right, $\delta = -0.18$).	35
3.2	Credible subgroups using stochastic search variable selection and model averaging (left) and using the most frequently selected model only (right).	37
3.3	Credible subgroups (80%) using stochastic search variable selection and model averaging (left) and using the most frequently selected model only (right).	38

3.4	Estimated nonlinear effects and 95% pointwise credible bands on standardized covariate scale, relative to sample mean. Rug plot represents observed covariate distribution.	44
3.5	Credible subgroups at the 95% level. Green points represent the exclusive credible subgroup, and yellow the remainder of the inclusive credible subgroup.	45
3.6	Credible subgroups at the 95% level for the linear (left) and BART (right) models.	46
3.7	Posterior mean PTE surfaces for the semiparametric (left) and linear (right) models.	46
4.1	Simulated sensitivity (left column) and specificity (right column) for a study with $A = 2$ arms and varying number of endpoints (top row) and a study with $K = 1$ endpoints and a varying number of arms (bottom row). In most cases, sensitivity falls and specificity remains high as more arms or endpoints are added.	57
4.2	Credible subgroups for individual endpoint-competitor combinations at the 50% level.	61
4.3	Credible subgroups for individual endpoint-competitor combinations at the 50% level.	62
5.1	The shaded region is the region for which the power to detect a 1 standard deviation (5-point) benefit at the 95% credible level is at least 5% when the entire empirical covariate space is used. Patients receiving placebo are represented by \times , and those receiving the standard of care by $+$. . .	69
5.2	Direct and mean-aligned conditional bootstrap sample of $\widehat{W}_{\alpha,C}^*$ (left) and direct and conditional approximations of $\widehat{L}(z)$ (right).	72
5.3	Uncertainty in credible subgroups. Orange points are part of the estimated exclusive credible subgroup but not part of the 0.5th percentile of exclusive credible subgroups, and blue points are not part of the estimated exclusive credible subgroup but are part of the 99.5th percentile of exclusive credible subgroups.	73

5.4	Funnel plots for two previously analyzed datasets. Green points are members of the exclusive credible subgroup, and yellow points are part members of the inclusive but not exclusive credible subgroup.	75
5.5	Example of calculator for reporting credible subgroup results, using the results of the multi-trial dataset analysis.	77

Chapter 1

Identifying the Benefiting Subgroup

1.1 Addressing treatment effect heterogeneity

It is well known that different patients suffering from a given ailment will have different outcomes. In fact, this understanding is central to the usefulness of statistics in the health sciences, for if it were not the case, then we could determine the effect of an intervention by applying it to a single patient and withholding it from another. The earliest clinical trials addressed this variability in outcomes by randomly assigning subjects to one of two groups, one to which the treatment is applied and another from which it is withheld, and comparing the average outcomes between the two groups. The difference between the observed mean outcomes is then an estimate of the *average treatment effect* (ATE) $\Delta \equiv E[Y|t = 1] - E[Y|t = 0]$, where Y denotes a numerical summary of the outcome and $t = 0, 1$ indicates the test treatment being withheld or applied, respectively. The ATE can then be interpreted as the expected difference in outcomes for a randomly selected patient if the treatment were to be given to versus withheld from them.

It has also long been recognized that some of the variation in outcomes can be attributed to *prognostic covariates*, patient characteristics observable before treatment that provide information about how those patients will fare regardless of treatment

choice. Thus linear regression becomes useful through models of the form¹

$$E[Y|\mathbf{x}; t] = \mathbf{x}^\top \boldsymbol{\beta} + t\gamma, \quad (1.1)$$

where \mathbf{x} is a vector of prognostic covariates, $\boldsymbol{\beta}$ is a vector of prognostic effect parameters, and $\gamma = \Delta$ is the average treatment effect. Such a model can yield substantially more precise estimates of the treatment effect if the included covariates are in fact correlated with outcomes.

Given that variation exists in patient outcomes and can be partially explained by baseline characteristics, it does not require a great leap of the imagination to suspect that the same might hold for the treatment effect, i.e., that there is *treatment effect heterogeneity* with respect to those characteristics, which we term *predictive covariates*. Prognostic and predictive covariates (which may overlap) then enter the regression model as main effects and interaction effects with the treatment, respectively:

$$E[Y|\mathbf{x}, \mathbf{z}; t] = \mathbf{x}^\top \boldsymbol{\beta} + t\mathbf{z}^\top \boldsymbol{\gamma}, \quad (1.2)$$

where \mathbf{z} is a vector of predictive covariates and $\boldsymbol{\gamma}$ is a vector of predictive effect parameters. Now, rather than relying on the average treatment effect, the *personalized treatment effect* (PTE) may be defined as $\Delta(\mathbf{z}) \equiv E[Y|\mathbf{x}, \mathbf{z}; t = 1] - E[Y|\mathbf{x}, \mathbf{z}; t = 0]$. In the case of (1.2), $\Delta(\mathbf{z}) = \mathbf{z}^\top \boldsymbol{\gamma}$. The PTE has the more clinically useful interpretation of the expected difference in outcomes for a randomly selected patient *with specific baseline characteristics* \mathbf{z} , if the treatment were given to versus withheld from them.

Acknowledgment of treatment effect heterogeneity has led to increased interest in *precision medicine* (also called *personalized medicine*), an approach to treatment that takes into account individual patients' characteristics such as demographics, genetics, lifestyle, environment, and more. Of course, in order for practitioners to take individual patients' characteristics into account for individualized care, information on how those characteristics determine the treatment effect must exist in a useful and accessible form.

The first step in providing practitioners with the information necessary to personalize care is to collect data that can be used to produce it. Specifically, studies must enroll a diverse cohort of subjects. To this end, the 2012 Food and Drug Administration Safety

¹ The variable t is separated from \mathbf{x} with a semicolon to indicate that it is set by the experimenter, as opposed to being observed as a characteristic of the patient.

and Innovation Act (FDASIA) Section 907, Reporting of Inclusion of Demographic Subgroups in Clinical Trials and Data Analysis in Applications for Drugs, Biologics, and Devices, required the FDA to publish a report “addressing the extent to which clinical trial participation and the inclusion of safety and effectiveness data by demographic subgroups including sex, age, race, and ethnicity, is included in applications submitted to the Food and Drug Administration.” [1] As part of the report, the FDA presented the FDA Action Plan to Enhance the Collection and Availability of Demographic Subgroup Data, which aimed to improve the completeness and quality of demographic subgroup data, to identify barriers to subgroup enrollment in clinical trials and employ strategies to encourage greater participation, and to make demographic subgroup data more available and transparent [2]. Although the act and subsequent report only address demographic factors and leave out genetic, lifestyle, and environmental ones, demographic factors are a natural starting point due to their relative ease of collection, public accessibility, and potential to partially address health disparities. To expand the range of studied characteristics, President Obama in 2015 announced the Precision Medicine Initiative [3], which in part aims to expand the use of genetically-based clinical trials for cancer treatments and form a national research cohort of over one million American volunteers. The cohort will contribute data including “medical records; profiles of the patient’s genes, metabolites, and microorganisms in and on the body; environmental and lifestyle data; patient-generated information; and personal device and sensor data.” Finally, the 21st Century Cures Act [4] passed in 2016 seeks to streamline clinical trials, including by means of Bayesian methods, and focuses on cancer and Alzheimer’s disease, areas of intense interest within personalized medicine.

1.2 Modes of inference for personalized medicine

Once data usable for investigating treatment effect heterogeneity is available, we may choose from several inferential goals, depending largely how many patients and how many interventions are under consideration, as well as whether the focus is on screening or optimization. In all of these contexts, many quantities are estimated (the treatment effect for each combination of type of patient and intervention), and as a result many hypotheses are tested (whether or not there is a benefit under such a combination).

Thus the idea of multiple testing is often relevant.

The *multiple testing problem* refers to the statistical phenomenon that when many questions are asked and answered, each with a small probability of being wrong, individually, the probability of producing at least one wrong answer (the *familywise error rate*) may still be quite large. In the context of personalized medicine, a *Type I error* usually refers to identifying a type of patient as benefiting from an intervention when in fact they do not, while a *Type II error* refers to failing to recognize that a type of patient benefits from an intervention. In the same context, we take *screening* to mean determining whether a given patient-intervention combination is acceptable in some sense, and *optimization* to mean determining the best intervention to apply to each patient, or the patients for which a given intervention is most effective. Screening usually requires careful attention to multiple testing considerations, while optimization often does not.

1.2.1 One patient, many interventions

The most familiar case of one patient, many interventions is likely the patient-practitioner interaction. In this situation, a patient presents an ailment to a practitioner (e.g., primary care physician), and the practitioner must consider several possible interventions (e.g., the set of all drugs approved to treat a presented illness). While the practitioner may draw on his or her experiences with other patients as well as published information, the practitioner makes decisions regarding only the current patient. However, the practitioner must make decisions regarding many interventions.

In the screening problem, the practitioner must identify which among the available interventions would benefit the patient. In the pharmaceutical context, this problem is not often treated statistically, since ideally comparing the patient against the guidelines on the label would be sufficient to determine whether the treatment is acceptable according to regulators.

The optimization problem, in which the patient may leave it up to the practitioner to recommend the best treatment, has been studied in much greater detail. In particular, the field of dynamic treatment regimes (DTRs) [5] seeks methods to identify the optimal processes in which a sequence of interventions are applied to a single patient, taking into account that patient's baseline characteristics as well as responses to previous treatments. For example, a simple process for treating patients with hypertension

would be to prescribe drug A to males and drug B to females, try the opposite treatment if the first does not lower blood pressure sufficiently, and finally return to the first treatment if the second performs even worse. In most approaches to the optimization problem, the procedure is unbiased in the sense that the most effective treatment is most likely to be identified as such. However, unadjusted estimates of the magnitude of the effect of that treatment are biased upward due to selection bias.

1.2.2 Many patients, one intervention

Situations in which only one intervention is under consideration occur commonly during the treatment development, evaluation, and regulatory approval processes. For example, once a potentially therapeutic compound has been identified, it may be compared exclusively to the present standard of care for the rest of its development cycle (and in this case the experimental treatment is the only one being evaluated, as long as the standard of care is well understood). In these processes, it is the responsibility of regulators to ensure that patients for whom a treatment is approved benefit from the treatment. This places the burden of proof for safety and efficacy on the treatment developers, though in the past it seems that the required proof has been that the broad population benefits on average. However, assuming that regulators are proficient in screening proposed treatments, it may also be in the best interests of developers to identify early the types of patients who benefit, or benefit most, from an experimental treatment. Such foreknowledge could allow developers to focus their resources and efforts on populations in which they are more likely to succeed. Formally, we pose the benefiting subgroup identification problem as estimating $B \equiv \{z : \Delta(z) > 0\}$.

Strictly speaking, standard clinical trials generally consider many patients (or types of patients), but often make their primary conclusions by conflating these differing types in order to infer an average treatment effect in the entire population. Although average treatment effects can be used to infer that *some* subgroup of patients benefits from treatment, they can only be used to *identify* the benefiting subgroup (i.e., everyone or no-one) under the assumption that there is no heterogeneity sufficient to cause the personalized treatment effect to change sign. Such an assumption is implicit in the United States regulatory guidance. The FDA’s 1998 Guidance For Industry, E9 Statistical

Principles for Clinical Trials [6] (also called ICH-E9)² states in Section 5.7, Subgroups, Interactions, and Covariates,

The treatment effect itself may also vary with subgroup or covariate. For example, the effect may decrease with age or may be larger in a particular diagnostic category of subjects. In some cases such interactions are anticipated or are of particular interest (e.g., geriatrics); hence a subgroup analysis or a statistical model including interactions is part of the planned confirmatory analysis. In most cases, however, subgroup or interaction analyses are exploratory and should be clearly identified as such; they should explore the uniformity of any treatment effects found overall. In general, such analyses should proceed first through the addition of interaction terms to the statistical model in question, complemented by additional exploratory analysis within relevant subgroups of subjects, or within strata defined by the covariates. When exploratory, these analyses should be interpreted cautiously. Any conclusion of treatment efficacy (or lack thereof) or safety based solely on exploratory subgroup analyses is unlikely to be accepted.

Thus in most cases the detection of a positive ATE and lack of overwhelming evidence of heterogeneity is sufficient to warrant acceptance in a large population. Such skepticism of subgroup analyses is understandable, as they have frequently misled investigators due to the inherent multiple testing problem and associated inflated Type I error rate: as more subgroups are examined without adjustments for multiplicity, the likelihood of observing a transient difference in treatment effect approaches certainty.

In the hope of mitigating the propensity of subgroup analyses to produce false positives, it has been recommended to carry them out only after identifying effect heterogeneity by means of an interaction test [7]. This recommendation comes with the acknowledgment that in trials designed to detect an ATE, such interaction tests are underpowered. This is presented as a strength of the approach, with the argument that such interactions are uncommon, and that the lack of power in interaction tests accurately reflects the scant evidence of heterogeneity. Others [8] consider this to be a

² The guidance was developed at the International Conference of Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

deficiency and recommend data mining methods for variable selection and model building as exploratory techniques. We argue that such lack of power, and in fact the general strategy of assuming homogeneity in the absence of compelling evidence to the contrary, is backwards when attempting to identify the benefiting subgroup. As the burden of proof lies with the treatment developer, this strategy creates an incentive to design studies with enough power to identify an ATE but not treatment effect heterogeneity. It is relatively easy to achieve such a “sweet spot” due to the gap in power between ATE and interaction tests for a given sample size.

Instead, the treatment effect should not be assumed homogeneous without prior evidence, and when present, the assumption must be explicit. Statistically, developers should allow for the presence of heterogeneity and model accordingly. For example, it has been proposed to analyze the response variable via a Bayesian linear model with skeptical (informative and centered around zero) priors on treatment-covariate interaction parameters [9]. Alternatively, tree-based regression methods have been proposed. Structures related to classification and regression trees (e.g., CART, BART) [10, 11, 12] have advantages of straightforward “flowchart-style” interpretation often used by clinicians and the ability to capture complex and nonlinear relationships. Finally, concerns about Type I error inflation should be addressed formally rather than swept under the rug through ad hoc use of underpowered tests. Adaptive signature designs and extensions [13, 14] take an approach of constructing via general classification models a subgroup thought to benefit and then perform a single test for the average treatment effect in that subgroup, however the inferential statements available from such procedures do not directly correspond to identification of the benefiting subgroup. This dissertation focuses on a confirmatory approach to benefiting subgroup identification that allows treatment effect heterogeneity in the absence of explicit assumptions to the contrary, and directly addresses multiplicity concerns.

Substantial attention has also been given to the various optimization problems for many patients and one treatment. Methods have been proposed that aim to partition the types of patients into two groups that show the greatest difference in treatment effect [15], and to search for subgroups with enhanced treatment effects relative to the general population [16, 17]. In these cases, multiplicity corrections are not as critical because they are meant to be used as exploratory tools rather than confirmatory.

1.2.3 Many patients, many interventions

The case of many patients, many interventions arises most naturally in the context of policymaking by entities intermediate to regulators and practitioners, e.g., hospitals, payers, and practitioner associations.

The screening problem often retains elements of both the one patient, many interventions and many patients, one intervention cases. In particular, multiplicity control may be important both in terms of types of patients and different interventions: a Type I error is incorrectly declaring that some patient-treatment combination is beneficial. However, the details of the multiplicity correction for multiple treatments may be flexible, especially when treatments have already been approved as generally safe and effective by regulators and each treatment has its own study arm, so that the notion that the comparisons all come from the same trial arises solely from a common control arm and, e.g., common organizational infrastructure. We consider this case in more detail in Chapter 4.

Recent work has also addressed the optimization problem, especially in the context of clinical trials for developing treatment policy. For example, the SUBA design [18] provides a tree-based algorithm for constructing subgroups and allocating patients adaptively to the best subgroup-specific treatments. Additionally, methods in network meta analysis [19] often seek to identify the best among a variety of treatments already on the market, sometimes acknowledging treatment effect heterogeneity [20].

1.2.4 Frequentist or Bayesian?

Because this dissertation deals primarily with interactions between developers and regulators, inference must ultimately be frequentist in that the familywise Type I error rate must be controlled. However, many aspects of the Bayesian inferential framework prove to be both natural and expedient. Bayesian hierarchical models provide an intuitive conceptualization of effect heterogeneity, and careful choice of priors on treatment-covariate interaction effects can be used to tune the level of heterogeneity of the treatment effect among different types of patients to either reflect prior information or elicit desired operating characteristics. Additionally, Bayesian computation schemes which provide Monte Carlo samples from the joint posterior distribution of the PTEs

substantially simplify methods for multiple comparisons. Thus our approach will be to model the treatment effect in a Bayesian framework but require certain frequentist properties, achieved by asymptotic correspondence between the parameter posterior and estimator sampling distributions, and verified via simulation studies.

In addition to the familywise Type I error, we will pay special attention to the average sensitivity and specificity of estimators for the benefiting subgroup, which play roles analogous to power and Type I error control. For an estimate \widehat{B} of the benefiting subgroup B and a measure μ_Z on the predictive covariate space, we define the sensitivity and specificity of \widehat{B} as $\mu_Z(\widehat{B} \cap B) / \mu_Z(B)$ and $\mu_Z(\widehat{B}^c \cap B^c) / \mu_Z(B^c)$, respectively, and take the frequentist expectation of these quantities over the distribution of \widehat{B} . We take the measure μ_Z to be uniform over some restriction of the covariate space, though empirical estimates of the population covariate density may also be useful.

1.3 Clinical motivation and plan of dissertation

1.3.1 Alzheimer’s disease

Our motivating datasets stem from several clinical trials of Alzheimer’s disease (AD) treatments carried out by AbbVie, Inc. While effective treatment strategies for AD are in their infancies, a number of risk factors for the disease are known. For example, advanced age and the presence of the Apolipoprotein E4 (ApoE4) allele dramatically increase the risk of AD, while longer education and higher intelligence appear somewhat protective [21].

One of the current standard of care (SOC) treatments for AD is Donepezil (trade name Aricept), a palliative medication approved in the United States for the treatment of mild to severe dementia resulting from AD [22]. It is not thought to alter the course or progression of AD itself, and sometimes causes nausea, diarrhea, and vomiting [23]. Donepezil was the SOC used in all of the trials described below.

In some trials, a test treatment was compared to both the SOC and placebo. The primary endpoint was the 12- or 24-week improvement in cognitive function as measured by the eleven-point Alzheimer’s Disease Assessment Scale–Cognitive Subscale (ADAS-Cog 11) [24]. Here improvement is defined as the severity score at baseline minus the severity score at the individual ends of follow-up. The test treatments under study in

these trials were abandoned before completion of the regulatory process, and consequently the full details of the trials are not publicly available.

1.3.2 Simple add-on therapy dataset

The first dataset is from a trial of a compound we refer to as ATT-1 at three dose levels. The compound was evaluated as an add-on therapy to the SOC, and evaluated to the SOC plus a placebo. We present a comparison of the low-dose test treatment to the placebo on a subset of patients of the sponsor’s interest. There are 41 such patients, 25 receiving the placebo and 16 receiving the treatment. The primary endpoint is 12-week improvement, and the potential predictive covariates are the baseline severity on the ADAS-Cog 11 scale, age, sex, and carrier status of the ApoE4 biomarker. This dataset provides a simple, introductory example for an analysis using a standard Bayesian formulation of the multiple linear regression model in Chapter 2.

1.3.3 Multi-endpoint dataset

The second dataset is from a trial of three dose levels of ATT-1 as a monotherapy versus the SOC and placebo as separate control arms. Data from all five arms are presented, totaling 331 patients. Responses are available for both the efficacy endpoint (24-week ADAS-Cog 11 improvement) and the safety endpoint (reporting of at least one adverse event indicated by the physician to be possibly related to the treatment). The same predictive covariates are available as in the add-on therapy dataset. This dataset provides an example application of our approach to the generalized linear model in Chapter 3 and the multi-endpoint, multi-treatment setting in Chapter 4.

1.3.4 Multi-trial dataset

The final dataset is the combination of four clinical trials of different test treatments as monotherapies versus the SOC and placebo. The four trials were carried out in the same set of centers in short succession. In all four trials the primary endpoint is 12-week improvement, and the potential predictive covariates are baseline severity, sex, ApoE4 carrier status, and the rate of decline in cognitive function as measured by the Mini-Mental State Exam (MMSE) from the onset of first symptoms to study baseline.

We present data from the SOC and placebo arms totaling 369 patients to illustrate the application of our approach using semiparametric and nonparametric regression models in Chapter 3.

1.3.5 Implementation and future work

While Chapters 2–4 describe primarily theoretical work, later chapters address more practical considerations, including topics in implementation for clinical trials and implications for future research. In particular, Chapter 5 considers power computations, Monte Carlo precision, diagnostics, reporting of results, and software usage, while Chapter 6 concludes by summarizing the work contained in the dissertation, discussing its significance both in clinical trials and in a broader range of topics, and presenting possible avenues for future work.

Chapter 2

Credible Subgroups

This chapter introduces the foundational ideas of credible subgroups as an estimator for the benefiting subgroup. The case treated in Section 2.1 is that of a single, conditionally normally distributed response in a standard Bayesian formulation of the multiple linear regression model. Later chapters extend these ideas to more general regression techniques (Chapter 3) and inferential goals (Chapter 4), and provide additional tools for practical implementation of the approach (Chapter 5). Section 2.2 provides a simulation study of the operating characteristics of the credible subgroups approach, and Section 2.3 applies the method to the simple add-on therapy dataset.

We will use the descriptions “types of patients” and “covariate points” interchangeably, and use “benefit” to mean a personalized treatment effect (PTE) or conditional average treatment effect greater than some threshold δ , rather than a causal or potential outcomes conception of benefit for individuals.

2.1 Bounds as estimators

Recall the linear regression model

$$E[Y|\mathbf{x}, \mathbf{z}; t] = \mathbf{x}^T\boldsymbol{\beta} + t\mathbf{z}^T\boldsymbol{\gamma}, \tag{2.1}$$

where Y is the response, \mathbf{x} is a vector of prognostic covariates with corresponding parameter vector $\boldsymbol{\beta}$, \mathbf{z} is a vector of predictive covariates with corresponding parameter vector $\boldsymbol{\gamma}$, and t is the treatment indicator. Each of \mathbf{x} and \mathbf{z} may contain an intercept

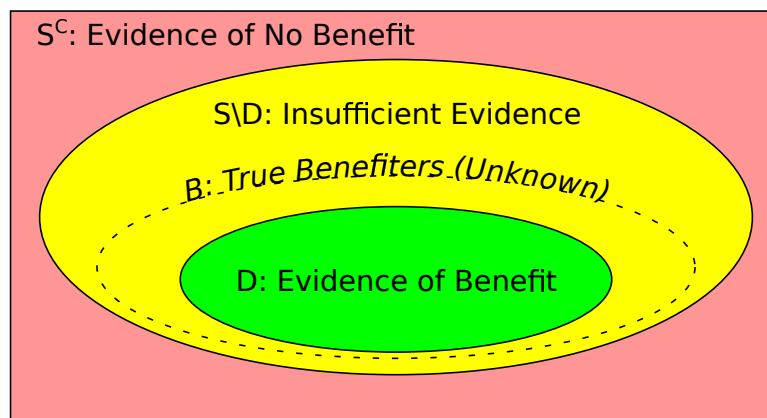


Figure 2.1: Interpretation of the trichotomy of the covariate space induced by the credible subgroup pair (D, S) relative to the true benefiting subgroup B .

term, and covariates may appear in both. The existing methods for estimating the benefiting subgroup B described in Section 1.2.2 provide a single estimate, \hat{B} of B that is meant to be a “best guess.” Even when \hat{B} is constructed based on a statistical test or bound, such as $\hat{B} = \{z : P[\Delta(z) > \delta | \mathbf{y}] \geq 1 - \alpha\}$, the single estimate is analogous to a point estimate of a univariate quantity. For such estimators the illustration of global uncertainty is not straightforward, as the basic inferences in such processes concern the inclusion or exclusion of covariate points from the benefiting subgroup individually. We instead propose a “bound” estimator analogous to credible (or confidence) intervals: a *credible subgroup pair* (D, S) ¹ for which D aims to contain *only* types of patients who benefit and S aims to contain *all* types of patients who benefit.

The pair (D, S) defines a trichotomy of the predictive covariate space, from which we may conclude that all patients in D have a treatment effect greater than a threshold δ , and that those in the complement S^c of S have treatment effect at most δ , while deferring conclusions about patients in the *uncertainty region* $S \setminus D$ (S remove D) until more information is available. This partition of the predictive covariate space is illustrated in Figure 2.1.

¹ The choice of symbols D and S arises from set-theoretical language describing intersections (German, *Durchschnitte*) and unions (French, *sommes*) of collections of sets, which are relevant to their construction in Section 2.1.2. The nomenclature may be most familiar from discussion of G_δ and F_σ sets.

Formally, we require a $1 - \alpha$ credible subgroup pair (D, S) to be such that

$$P(D \subseteq B \subseteq S | \mathbf{y}) \geq 1 - \alpha \quad (2.2)$$

in the Bayesian sense, with B as the random quantity. In most cases, we will additionally construct (D, S) so that $P(D \subseteq B \subseteq S) \geq 1 - \alpha$ in the frequentist sense, with (D, S) as the random quantity. We term D the *exclusive credible subgroup*, since the posterior probability that D contains *only* \mathbf{z} for which $\Delta(\mathbf{z}) > \delta$ is at least $1 - \alpha$. Similarly, we call S the *inclusive credible subgroup*, since the posterior probability that S contains *all* \mathbf{z} such that $\Delta(\mathbf{z}) > \delta$ is at least $1 - \alpha$. While there are many ways of satisfying these two conditions, taking the credible subgroups $D \equiv \{\mathbf{z} : P[\Delta(\mathbf{z}) > \delta | \mathbf{y}] > 1 - \alpha'/2\}$ and $S \equiv \{\mathbf{z} : P[\Delta(\mathbf{z}) > \delta | \mathbf{y}] > 1 - \alpha'/2\}$ is intuitive and yields unique pairs up to specification of $\alpha' \leq \alpha$. The two-sided threshold $\alpha'/2$ is used here because we will construct our credible subgroups using symmetric simultaneous confidence bands: roughly, covariate points for which the lower bound is greater than δ are included in D and those for which the upper bound is greater than δ are included in S . We discuss three methods for choosing α' .

First, for some level $\alpha \in (0, 1)$, let $G_{\alpha, \mathbf{y}}$ be the $1 - \alpha$ highest posterior density credible region for the interaction parameters $\boldsymbol{\gamma} | \mathbf{y}$. To every element $\hat{\boldsymbol{\gamma}} \in G_{\alpha, \mathbf{y}}$ there corresponds a half-space $B_{\hat{\boldsymbol{\gamma}}}$ of the predictive covariate space with $\hat{\Delta}(\mathbf{z}) \equiv \mathbf{z}^\top \hat{\boldsymbol{\gamma}} > \delta$ for all $\mathbf{z} \in B_{\hat{\boldsymbol{\gamma}}}$. Let \mathcal{B} be the collection of all $B_{\hat{\boldsymbol{\gamma}}}$ corresponding to $\hat{\boldsymbol{\gamma}} \in G_{\alpha, \mathbf{y}}$. Let D and S be the intersection and union, respectively, of all member sets of \mathcal{B} . Then (2.2) is satisfied. We further describe this *highest posterior density* (HPD) method of finding credible subgroups in Section 2.1.2.

The HPD method assumes that the entire covariate space is of interest, and is thus underpowered when only a subset of the covariate space is considered. Examples of restrictions include indicator variables that can only take values 0 or 1, and numerical covariates for which investigators are only concerned with values that lie within some finite range. The restriction of the entire unbounded covariate space to a bounded one can dramatically reduce the size of simultaneous credible bands for treatment effects, and thus the exclusive credible subgroup can often be expanded and the inclusive credible subgroup contracted. We discuss a *restricted covariate space* (RCS) procedure for handling these cases in Section 2.1.3.

The HPD and RCS methods take advantage of the fact that credible regions for the regression parameters asymptotically agree with the corresponding frequentist confidence regions under an uninformative prior. Thus not only is there at least $1 - \alpha$ posterior probability that $D \subseteq B \subseteq S$, but treating B as fixed, $1 - \alpha$ is an approximate lower bound on the frequency with which $D \subseteq B \subseteq S$, often a desirable frequentist property. When such a frequentist property is not necessary, a larger exclusive credible subgroup and a smaller inclusive credible subgroup may be obtained for which the posterior probability that $D \subseteq B \subseteq S$ is closer to $1 - \alpha$. We discuss such a *purely Bayesian* (PB) approach in Section 2.1.4.

2.1.1 A normal hierarchical linear model

We now review a normal hierarchical linear model setting for which we will develop examples of our benefiting subgroup selection tools. Let $\phi = (\beta, \gamma)$ be the combined vector of effect parameters. For each patient i , let Y_i be the response, \mathbf{x}_i be the prognostic covariate vector, \mathbf{z}_i be the predictive covariate vector, and $t_i \in \{0, 1\}$ indicate assignment to the control or treatment arm, respectively. Let \mathbf{X} be the $n \times p$ prognostic design matrix with the \mathbf{x}_i^\top as rows, \mathbf{Z} be the $n \times q$ predictive design matrix with the \mathbf{z}_i^\top as rows, and \mathbf{T} be the $n \times n$ diagonal treatment matrix $\text{diag}(t_1, \dots, t_n)$. Consider the model

$$\begin{aligned} \mathbf{Y} | \mathbf{X}, \mathbf{Z}, \mathbf{T}, \beta, \gamma, \sigma^2 &\sim \text{Normal} [\mathbf{X}\beta + \mathbf{TZ}\gamma, \sigma^2 \mathbf{\Sigma}], \\ \phi | \sigma^2 &\sim \text{Normal} [\boldsymbol{\nu}, \sigma^2 \mathbf{R}], \\ \sigma^2 &\sim \text{InverseGamma} [a_0, b_0], \end{aligned} \tag{2.3}$$

where $\mathbf{\Sigma}$, $\boldsymbol{\nu}$, \mathbf{R} , a_0 , and b_0 are hyperparameters assumed known. The variance σ^2 is included in the prior scale for ϕ for conjugacy. With $\mathbf{W} = (\mathbf{X} \ \mathbf{TZ})$ as the full design matrix, the first line of (2.3) becomes $\mathbf{Y} | \mathbf{W}, \phi, \sigma^2 \sim \text{Normal} [\mathbf{W}\phi, \sigma^2 \mathbf{\Sigma}]$.

The posterior distribution of ϕ conditioned on σ^2 is then [25]

$$\begin{aligned} \phi | \mathbf{y}, \mathbf{W}, \sigma^2 &\sim \text{Normal} [\mathbf{H}_\phi \mathbf{h}_\phi, \sigma^2 \mathbf{H}_\phi], \\ \mathbf{H}_\phi^{-1} &= \mathbf{W}^\top \mathbf{\Sigma}^{-1} \mathbf{W} + \mathbf{R}^{-1}, \\ \mathbf{h}_\phi &= \mathbf{W}^\top \mathbf{\Sigma}^{-1} \mathbf{y} + \mathbf{R}^{-1} \boldsymbol{\nu}, \end{aligned} \tag{2.4}$$

and the posterior distribution of σ^2 is

$$\begin{aligned}\sigma^2|\mathbf{y}, \mathbf{W} &\sim \text{InverseGamma}[a, b], \\ a &= a_0 + \frac{n}{2}, \\ b &= b_0 + \frac{1}{2} \left(\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} + \boldsymbol{\nu}^\top \mathbf{R}^{-1} \boldsymbol{\nu} - \mathbf{h}_\phi^\top \mathbf{H}_\phi \mathbf{h}_\phi \right).\end{aligned}\tag{2.5}$$

Thus the marginal posterior of $\boldsymbol{\phi}$ is the multivariate Student's t distribution

$$\boldsymbol{\phi}|\mathbf{y}, \mathbf{W} \sim \text{Student} \left[\mathbf{H}_\phi \mathbf{h}_\phi, \frac{b}{a} \mathbf{H}_\phi \right],\tag{2.6}$$

and the marginal posterior of $\boldsymbol{\gamma}$ is

$$\boldsymbol{\gamma}|\mathbf{y}, \mathbf{W} \sim \text{Student} \left[\mathbf{H} \mathbf{h}, \frac{b}{a} \mathbf{H} \right],\tag{2.7}$$

where \mathbf{H} is the submatrix of \mathbf{H}_ϕ and $\mathbf{H} \mathbf{h} = \bar{\boldsymbol{\gamma}}$ is the subvector of $\mathbf{H}_\phi \mathbf{h}_\phi$ corresponding to the coordinates of $\boldsymbol{\gamma}$ only.

2.1.2 Highest posterior density credible subgroups

Let $G_{\alpha, \mathbf{y}}$ be the highest posterior density (HPD) $1 - \alpha$ credible set for $\boldsymbol{\gamma}$. We define D to be the intersection all sets $B_\gamma \equiv \{\mathbf{z} : \mathbf{z}^\top \boldsymbol{\gamma} > \delta\}$ for all $\boldsymbol{\gamma} \in G_{\alpha, \mathbf{y}}$, and S to be the union of all such sets. Equivalently, a given predictive covariate vector \mathbf{z} is in D if and only if $\mathbf{z}^\top \boldsymbol{\gamma} > \delta$ for *all* $\boldsymbol{\gamma} \in G_{\alpha, \mathbf{y}}$, and in S if and only if $\mathbf{z}^\top \boldsymbol{\gamma} > \delta$ for *any* $\boldsymbol{\gamma} \in G_{\alpha, \mathbf{y}}$.

Under the marginal posterior distribution (2.7), the boundary of $G_{\alpha, \boldsymbol{\gamma}}$ is the ellipsoid

$$(\boldsymbol{\gamma} - \bar{\boldsymbol{\gamma}})^\top \left(\frac{b}{a} \mathbf{H} \right)^{-1} (\boldsymbol{\gamma} - \bar{\boldsymbol{\gamma}}) = qF(1 - \alpha, q, 2a),\tag{2.8}$$

where $F(1 - \alpha, q, 2a)$ is the $1 - \alpha$ quantile of the F distribution on q numerator and $2a$ denominator degrees of freedom. It can be shown [26] that for a given \mathbf{z} ,

$$\mathbf{z}^\top \boldsymbol{\gamma} \in \mathbf{z}^\top \bar{\boldsymbol{\gamma}} \pm \sqrt{qF(1 - \alpha, q, 2a)} \sqrt{\mathbf{z}^\top \left(\frac{b}{a} \mathbf{H} \right) \mathbf{z}}\tag{2.9}$$

for precisely the $\boldsymbol{\gamma} \in G_{\alpha, \mathbf{y}}$. Thus D is the set of \mathbf{z} for which the lower bound of (2.9) is greater than δ , and S is comprised of those for which the upper bound is at least δ . These bounds correspond to the Scheffé simultaneous confidence bands for the frequentist normal linear model, though with a slightly underestimated variance parameter due to its denominator of n versus $n - 1$. The formulation in terms of simultaneous credible bands will be the basis for subsequent constructions of credible subgroups.

2.1.3 Restricted covariate spaces

The HPD method uses bands that are exact when \mathbf{z} ranges over all of \mathbb{R}^q and is conservative when only a subset C of the covariate space is of interest. In such cases, the narrower band

$$\mathbf{z}^\top \boldsymbol{\gamma} \in \mathbf{z}^\top \bar{\boldsymbol{\gamma}} \pm W_{\alpha, C}^* \sqrt{\mathbf{z}^\top \left(\frac{b}{a} \mathbf{H} \right) \mathbf{z}} \quad (2.10)$$

may be used in the same manner [27], where $W_{\alpha, C}^*$ is the $1 - \alpha$ quantile of the distribution of

$$W_C = \sup_{\mathbf{z} \in C} \frac{|\mathbf{z}^\top (\boldsymbol{\gamma} - \bar{\boldsymbol{\gamma}})|}{\sqrt{\mathbf{z}^\top \left(\frac{b}{a} \mathbf{H} \right) \mathbf{z}}}. \quad (2.11)$$

The distribution of W_C is usually analytically intractable, but $W_{\alpha, C}^*$ may be estimated via Monte Carlo methods by drawing a sample from the posterior (2.7) of $\boldsymbol{\gamma}$ and computing the corresponding values of W_C . When continuous covariates are present, a grid may be used for approximation. This restricted-space approach to constructing simultaneous credible bands will be our choice for most extensions to the credible subgroups method.

2.1.4 Purely Bayesian credible subgroups

The HPD and RCS methods leverage the frequentist properties of estimates of parameters and linear combinations of those parameters to make frequentist coverage guarantees, but are conservative in terms of posterior probabilities only. Exact credible subgroups may be obtained by replacing $\sqrt{qF(1 - \alpha, q, 2a)}$ in (2.9) with some smaller value r . This yields a larger exclusive credible subgroup and a smaller inclusive credible subgroup.

Given a sample from the posterior of $\boldsymbol{\gamma}$ and a finite set C of points in the predictive covariate space, Algorithm 1 provides a Monte Carlo method for estimating an appropriate value of r via binary search to yield $P(D \subseteq B \subseteq S)$ within some margin ϵ of $1 - \alpha$: When the set C or the posterior sample size are small, the algorithm may not reach the target precision for \hat{p} , in which case the smallest $\hat{p} > 1 - \alpha$ may be used.

Algorithm 1 Pure Bayes credible subgroup construction

- 1 Set search bounds $r_L = 0$ and $r_U = qF(1 - \alpha, q, 2a)$;
 - 2 **repeat**
 - 3 Set the working value for r to $\hat{R} = (r_L + r_U)/2$;
 - 4 Substitute \hat{r} for $\sqrt{qF(1 - \alpha, q, 2a)}$ in (2.9) to produce a working credible subgroup pair (\hat{D}, \hat{S}) ;
 - 5 Use the posterior sample of γ to produce a sample of B and estimate $\hat{p} = P(\hat{D} \subseteq B \subseteq \hat{S})$;
 - 6 If $\hat{p} > 1 - \alpha$ set $r_U = \hat{r}$, and if $\hat{p} < 1 - \alpha$ set $r_L = \hat{r}$;
 - 7 **until** \hat{p} is in $[1 - \alpha, 1 - \alpha + \epsilon]$
 - 8 Set $r = \hat{r}$.
-

2.1.5 Asymptotic properties of credible subgroups

The highest posterior density (HPD) regions and restricted covariate space (RCS) simultaneous credible bands share asymptotic properties with the corresponding Wald confidence regions and simultaneous confidence bands, respectively, based on maximum likelihood estimates. For the parametric model with parameter vector $\boldsymbol{\theta}$, we know that under certain regularity conditions, as the number n of independent and identically distributed observations grows,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0) \xrightarrow{d} \text{Normal}[0, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}], \quad (2.12)$$

where $\mathbf{J}(\boldsymbol{\theta}_0)$ is the Fisher information matrix. Similarly [28], the transformed posterior

$$\sqrt{n}([\boldsymbol{\theta}|\mathbf{y}] - \boldsymbol{\theta}_0) \xrightarrow{d} \text{Normal}[0, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}]. \quad (2.13)$$

Thus under the regularity conditions for (2.12) and (2.13), the HPD regions and RCS simultaneous credible bands (2.10) may also be interpreted as Wald confidence regions and simultaneous confidence bands over the same domain, allowing frequentist asymptotic interpretations of the credible subgroups so derived. The pure Bayes (PB) method fails to have a similar frequentist interpretation because it is based on the relationship

not of well-behaved estimates to parameters, but of \widehat{D} and \widehat{S} to B , which is dependent on the true value of B .

2.1.6 A step-down multiple testing procedure

The construction of credible subgroups by the single-step multiple testing procedure comparing RCS credible bands to a threshold may be improved upon by a sequential, step-down testing procedure similar to the Holm-Bonferroni step-down procedure [29], which is well-known in the multiple hypothesis testing literature. For a set of M hypotheses, the Holm-Bonferroni procedure first tests all hypotheses using an M -way Bonferroni correction, and if the hypothesis with the lowest adjusted p -value [30] can be rejected, proceeds to test the remaining $M - 1$ hypotheses using an $(M - 1)$ -way Bonferroni correction, and so on. The procedure may also be specified to reject all hypotheses with sufficiently low adjusted p -values at each step.

An analog for constructing credible subgroups follows the same strategy, replacing the Bonferroni-corrected testing step with one comparing the bounds of the RCS credible bands to the threshold for benefit. Let C be a subset of interest of the covariate space, $\boldsymbol{\theta}$ be the vector of all model parameters (or, later, PTEs for nonparametric models), and $H_{\mathbf{z}} = \{\boldsymbol{\theta} : \Delta(\mathbf{z}) = \delta\}$. Then Algorithm 2 controls the overall Type I error rate at level α .

Algorithm 2 Step-Down Procedure

- 1 Let $M = 1$, $T_0 = C$, and $R_0 = \emptyset$ be the starting iteration, base test set, and base rejection set, respectively;
 - 2 **repeat**
 - 3 Let $T_M = T_{M-1} \setminus R_{M-1} = C \setminus (\bigcup_{m < M} R_m)$ be the new test set;
 - 4 Construct the two-sided $1 - \alpha$ restricted covariate space simultaneous confidence band (2.10) for $\Delta(\mathbf{z})$ over all $\mathbf{z} \in T_M$;
 - 5 Let R_M be the set of \mathbf{z} for which the band does not contain δ ;
 - 6 Increment M ;
 - 7 **until** $R_M = \emptyset$;
 - 8 Reject $H_{\mathbf{z}}$ for all $\mathbf{z} \in \bigcup_{m < M} R_m$.
-

The proof of validity for this procedure relies on showing that it is a closed testing procedure [31], in part by noting that $W_{\alpha, V}^* \leq W_{\alpha, U}^*$ for $V \subset U$. The full proof is available as the proof to Theorem 4 in Appendix A.2. If $H_{\mathbf{z}}$ is rejected, we may place \mathbf{z} in D when the posterior mean $\bar{\Delta}(\mathbf{z}) > \delta$ or S^c when $\bar{\Delta}(\mathbf{z}) < 0$, and if $H_{\mathbf{z}}$ is not rejected, we leave \mathbf{z} in $S \setminus D$.

2.1.7 Maximum credible levels

In general, *adjusted p-values* are computed such that if individual hypotheses are rejected if and only if the adjusted p -value for that hypothesis is less than α , then the familywise Type I error rate is controlled at the level α [30]. A Bayesian counterpart in the context of credible subgroups is the *maximum credible level* for a point \mathbf{z} —the highest credible level at which that covariate point is either included in D or excluded from S . Uses for maximum credible levels include communicating a more specific level of confidence for individual covariate points than a single credible level for the entire space, and allowing consumers of the statistical results to easily choose at which level they wish to construct credible subgroups while avoiding onerous additional computations.

For the single-step RCS credible subgroup procedure, the $1 - \alpha$ simultaneous credible

band at \mathbf{z} given by (2.10) may be written more generally as

$$\Delta(\mathbf{z}) \in \bar{\Delta}(\mathbf{z}) \pm W_{\alpha, C}^* \sqrt{\text{Var}[\Delta(\mathbf{z})]}. \quad (2.14)$$

Thus the highest credible level for which the band does not include δ at \mathbf{z} is

$$1 - \alpha_{\min} = \max \left\{ F_{W_C} \left(\pm \frac{\bar{\Delta}(\mathbf{z}) - \delta}{\sqrt{\text{Var}[\Delta(\mathbf{z})]}} \right) \right\}, \quad (2.15)$$

where F_{W_C} is the cumulative distribution function of W_C , estimated from the joint posterior of the $\Delta(\mathbf{z})$ in the same way as $W_{\alpha, C}^*$.

When computing maximum credible levels for the step-down procedure, the basic idea is to iterate as in Algorithm 2 and adjust (2.15) to account for the fact that the maximum credible level at a covariate point cannot exceed that of the covariate point in the previous iteration. Algorithm 3 produces for each $\mathbf{z} \in C$ the maximum credible level $l_{\mathbf{z}}$ at which the hypothesis that $\Delta(\mathbf{z}) = \delta$ is rejected, and may be used to quickly construct credible subgroups at various credible levels after performing a single expensive computation. The credible subgroups for any level $1 - \alpha$ are then $D = \{\mathbf{z} : \bar{\Delta}(\mathbf{z}) > \delta, l_{\mathbf{z}} \geq 1 - \alpha\}$ and $S^G = \{\mathbf{z} : \bar{\Delta}(\mathbf{z}) < \delta, l_{\mathbf{z}} \geq 1 - \alpha\}$.

Algorithm 3 Maximum Credible Levels

- 1 Let $M = 1$, $T_0 = C$, and $\mathbf{r}_0 = \emptyset$, similar to Algorithm 2;
 - 2 **repeat**
 - 3 Let $T_M = T_{M-1} \setminus \mathbf{r}_{M-1} = C \setminus (\bigcup_{m < M} \mathbf{r}_m)$ be the new test set;
 - 4 Compute a sample of the W_{T_M} as in equation (2.11) (only the draws for which the absolute Z -score was maximized at \mathbf{r}_{M-1} change, allowing a computational shortcut);
 - 5 Compute $q_i = \max \left\{ \hat{F}_{W_{T_M}} \left(\pm \bar{\Delta}(\mathbf{x}_i) / \sqrt{\text{Var}[\Delta(\mathbf{x}_i)]} \right) \right\}$ for $\mathbf{x}_i \in T_M$;
 - 6 Let $I = \arg \max_i q_i$, $\mathbf{r}_M = \mathbf{x}_I$, and record $l_I = \min \{q_I, q_{\mathbf{r}_{M-1}}\}$ as the maximum credible level for the test at \mathbf{r}_M ;
 - 7 Increment M ;
 - 8 **until** $M > |C|$.
-

2.2 Simulations

We perform a simulation study to evaluate certain frequentist properties of each method for constructing credible subgroup pairs. The property of primary interest is the frequency with which $D \subseteq B \subseteq S$ under a fixed parameter vector. We term this frequency as the *coverage rate*, in parallel with the usual coverage rate of, e.g., credible intervals.

We also wish to have a notion of the generalized width, or *size*, of the credible subgroup pair. A natural choice is to consider $\mu_Z(S \setminus D)$, where μ_Z is some measure on the covariate space. For example, if μ_Z is an estimate of the population covariate distribution, then the corresponding credible subgroup pair size is an estimate of the proportion of the population in the uncertainty region $S \setminus D$.

We are also able to treat each of the credible subgroups as a diagnostic test and compute sensitivities and specificities for D and S . These quantities measure how well the credible subgroups align with the benefiting subgroup. The sensitivity of D , $\mu_Z(D \cap B) / \mu_Z(B)$, is displayed here, and other related quantities in Appendix B.

In addition to comparing the three methods of constructing credible subgroups, we include in our simulations a nonsimultaneous method of identifying benefiting subgroups, which we call the *pointwise* method. The pointwise method uses the same normal linear model as the rest of this chapter, but does not account for multiplicity in constructing the credible subgroups; i.e., it sets $D = \{\mathbf{z} : P[\Delta(\mathbf{z}) > \delta | \mathbf{y}] \geq 1 - \alpha\}$.

We simulate 1000 datasets each containing $n = 40$ subjects to reflect the size of the example dataset used in the next section. Analogous results with $n = 100$ and $n = 350$ are presented in Appendix B. Each subject i has a covariate vector $\mathbf{x}_i = (1, x_{i2}, x_{i3})$ with $x_{i2} = 0, 1$ with equal probability and x_{i3} continuously uniformly distributed on $[-3, 3]$, a binary treatment indicator t_i taking values 0 and 1 with equal probability, and a normally distributed conditional response y_i . The covariates are used as both prognostic and predictive covariates and denoted \mathbf{x}_i and \mathbf{z}_i in the respective roles. The response has mean $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + t_i \mathbf{z}_i^\top \boldsymbol{\gamma}$ and variance $\sigma^2 = 1$. We fix $\boldsymbol{\beta} = \mathbf{0}$ and use six different values for $\boldsymbol{\gamma}$. We also present three simulations in which the effects of x_2 are nonlinear in order to evaluate the effects of misspecification. The “near-linear” configuration uses effects linear in $x'_2 = \sqrt{x_2 + 3} - \sqrt{3}$, “threshold” uses $x'_2 = x_2^{1/3}$, and “non-monotone” uses $x'_2 = 1/2 - (x_2/3)^2$.

We use a vague InverseGamma $[10^{-3}, 10^{-3}]$ prior for σ^2 and a Normal $[\mathbf{0}, \sigma^2 \mathbf{R}]$ prior on $\phi|\sigma^2$ with $\mathbf{R} = \text{diag}(10^4, 10^4, 10^4, 10^4, 1, 1)$, which is conservative with respect to interaction terms and vague with respect to other regression parameters. For each dataset, we compute credible subgroup pairs using each of the four methods at the 80% credible level (without the step-down procedure for the RCS method to retain comparability). To determine credible subgroups we use a grid search in which $z_1 = 1$, $z_2 = 0, 1$, and z_3 ranges from -3 to 3 in steps of 0.1 and include or exclude each covariate point on the grid from the subgroups as they satisfy or fail to satisfy the conditions specified in Section 2.1. Where a sample from the posterior of γ is needed, we use a sample of size 1000 drawn directly (not from a Markov chain). Finally, we also track how often an F test for treatment effect heterogeneity is significant at the 80% confidence level.

Table 2.1 displays the average summary statistics for 80% credible subgroup pairs under nine generating models ($n = 40$). Moving from the PB to RCS to HPD methods, coverage rate, pair size, and specificity of D increase, while sensitivity of D decreases. For both linear and nonlinear data generating mechanisms, the RCS and HPD methods have consistently conservative ($\geq 80\%$) coverage rate, while the PB method is sometimes conservative and at other times anticonservative.

Truth	Method	Coverage Rate	Pair Size	Sensitivity of D	Specificity of D	Heterog. Tests
$\gamma = (0, 0, 0)$	PB	0.46	0.75	–	0.87	0.18
	RCS	0.88	0.95	–	0.97	0.18
	HPD	0.91	0.97	–	0.98	0.18
	PW	0.43	0.59	–	0.79	0.18
$\gamma = (0, 0, 1)$	PB	0.82	0.25	0.76	0.99	1.00
	RCS	0.94	0.34	0.67	1.00	1.00
	HPD	0.96	0.38	0.64	1.00	1.00
	PW	0.46	0.13	0.87	0.98	1.00
$\gamma = (0, 1, 0)$	PB	0.55	0.55	0.68	0.83	0.45
	RCS	0.87	0.78	0.38	0.95	0.45
	HPD	0.91	0.82	0.33	0.96	0.45
	PW	0.47	0.39	0.79	0.71	0.45
$\gamma = (0, 1, 1)$	PB	0.77	0.25	0.81	0.99	1.00
	RCS	0.92	0.35	0.75	1.00	1.00
	HPD	0.95	0.38	0.72	1.00	1.00
	PW	0.41	0.14	0.89	0.97	1.00
$\gamma = (1, 0, 0)$	PB	0.99	0.25	0.75	–	0.18
	RCS	1.00	0.50	0.50	–	0.18
	HPD	1.00	0.56	0.44	–	0.18
	PW	0.97	0.13	0.87	–	0.18
$\gamma = (1, 1, 1)$	PB	0.73	0.24	0.87	0.97	1.00
	RCS	0.92	0.33	0.82	0.99	1.00
	HPD	0.94	0.35	0.80	0.99	1.00
	PW	0.43	0.15	0.92	0.93	1.00
Near-linear	PB	0.64	0.62	0.28	0.98	0.56
	RCS	0.92	0.84	0.13	1.00	0.56
	HPD	0.94	0.87	0.10	1.00	0.56
	PW	0.38	0.42	0.45	0.95	0.56
Threshold	PB	0.76	0.44	0.56	0.99	0.92
	RCS	0.93	0.61	0.40	1.00	0.92
	HPD	0.95	0.65	0.35	1.00	0.92
	PW	0.42	0.24	0.74	0.97	0.92
Non-monotone	PB	0.20	0.73	0.21	0.81	0.18
	RCS	0.80	0.93	0.06	0.95	0.18
	HPD	0.85	0.95	0.04	0.96	0.18
	PW	0.16	0.56	0.33	0.71	0.18

Table 2.1: Average summary statistics for 80% credible subgroup pairs as well as pointwise (PW) method (n=40). Statistics are averaged without undefined values, e.g. sensitivity of D when B is empty. Coverage rates at or above 80% and low pair sizes (analogous to interval lengths for interval estimation) are desired.

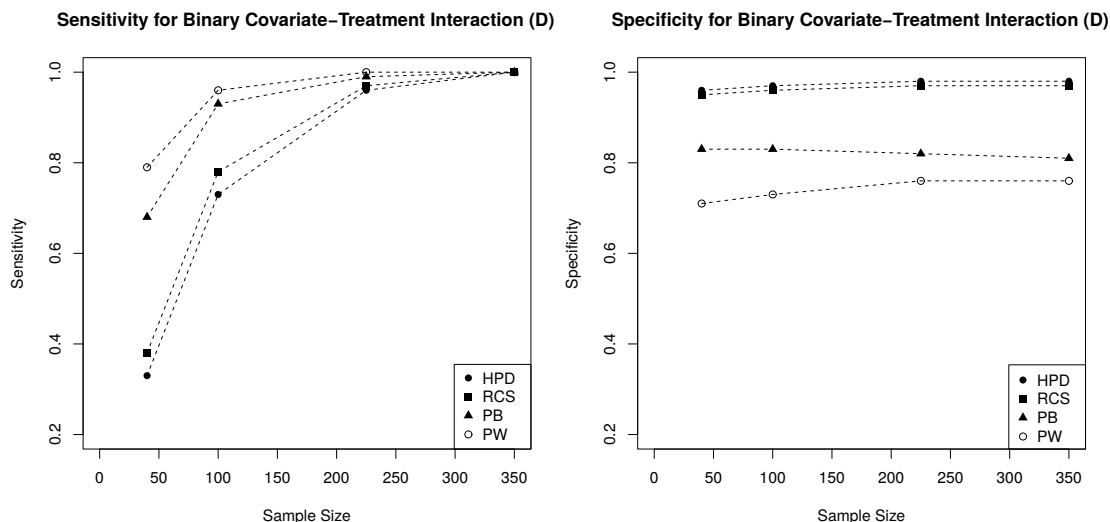


Figure 2.2: Diagnostic measure comparison in a case with a binary covariate-treatment interaction. Sensitivity (*left*) and specificity (*right*) of D in the case $\gamma = (0, 1, 0)$ (treatment effect is determined by a binary covariate). The multiplicity-correcting methods (HPD, RCS, and to a lesser extent PB) maintain extremely high specificity at the expense of sensitivity, especially for small sample sizes. Because the benefit is positive in one group and zero in its complement, the sensitivities of all methods approach 100% for large sample sizes while the specificities remain approximately constant.

The pointwise method yields generally tighter credible subgroups (smaller credible pair sizes) than the simultaneous methods, resulting in poorer coverage and specificity of D , but improved sensitivity of D . The primary advantage of the multiplicity-correcting methods is the extremely high specificity of D (and sensitivity of S), which are 100% whenever the coverage goal $D \subseteq B \subseteq S$ is met. However, the high specificity of D and sensitivity of S come at the price of lower sensitivity of D and specificity of S , especially for small samples. This trade-off may be favorable when extreme specificity is preferred over sensitivity (e.g., in a regulatory setting). Figure 2.2 illustrates the trade-off for D in the particularly interesting case of $\gamma = (0, 1, 0)$, a dichotomous predictive covariate for which one group has a constant positive benefit while the other has no benefit. Here, the PB method is nearly as sensitive as the uncorrected method, but only the fully corrected HPD and RCS methods deliver the extreme specificities desired by regulators.

Although the PB method is valid within a purely Bayesian context, we recommend against its use when strict frequentist guarantees are desired, and instead prefer the RCS or HPD methods. Further, we recommend the RCS method over the HPD method when the covariate space of interest is restricted, as the RCS method produces less conservative credible subgroup pairs and thus greater sensitivity of D . This advantage lessens as the covariate space becomes large and less discretized. In practical terms, the RCS method detects the most members within the benefiting population among methods that maintain the frequentist coverage guarantee. The step-down modification to the RCS method yields a 3–4% increase in the size of the exclusive credible subgroup over the displayed values under the linear data generating mechanisms. Though modest in payoff, this modification costs only additional computing time. Finally, the linearity assumption should be carefully considered, especially at larger sample sizes that can support the nonparametric models to be described in Chapter 3.

2.3 Analysis of simple add-on therapy dataset

We illustrate the credible subgroups approach on the add-on therapy dataset described in Section 1.3.2. We compare a low-dose treatment to a placebo on a subset of patients of the sponsor’s interest. There are 41 such patients, 25 receiving the placebo (`treatment = 0`) and 16 receiving the treatment (`treatment = 1`).

In addition to the intercept, four baseline measurements are of interest. The `severity` variable measures the progression of the disease at study baseline, so that high values indicate severe cognitive impairment. The `age` variable ranges from 58 to 90 at baseline, and `sex` is approximately 37% male (`sex = 1`) and 63% female (`sex = 0`). The `carrier` variable indicates the presence (`carrier = 1`) or absence (`carrier = 0`) of the ApoE4 allele, which 56% of the patients carry. The response of interest is `improvement`, defined as severity score at baseline minus end of follow up, so that a positive value of `improvement` indicates a positive outcome (decreased cognitive impairment). We assume that the responses are independent conditional on the covariates and that there is no heteroskedasticity ($\Sigma = \mathbf{I}$). We search for a population for which the personalized treatment effect $\Delta(\mathbf{z})$ is greater than zero for all members simultaneously at the 80% credible level ($\alpha = 0.20$).

We use all of the above baseline covariates as both prognostic and predictive variables. We also include the `sex:carrier` and `treatment:sex:carrier` interactions due to prior information that they may be important. The continuous covariates `severity` and `age` are standardized for computation and presentation of regression coefficients but are plotted in their original scales. An intercept and main treatment effect are also modeled.

Table 2.2 gives the posterior mean and standard deviation of effect parameters. Note that the overall treatment effect and only the interaction of treatment and age would be identified as significant at the 95% credible level with no multiplicity adjustment. The conclusion we wish to avoid is that the only treatment interaction is with age (see, e.g., `treatment:sex`). We consider this conclusion specious because a lack of evidence for strong interactions with sex, carrier status, and baseline severity does not imply a homogeneous treatment effect among levels of those covariates, and thus some patients may benefit from treatment while others may not. Instead, we wish to directly identify the baseline characteristics of patients for whom there is sufficient evidence of benefit from treatment, even when treatment–covariate interactions are weak.

We restrict our interest to the region of the covariate space where `severity` and `age` are within the ranges observed in the study participants, and proceed with the RCS method of identifying credible subgroups, including the step-down procedure. In order to estimate $W_{\alpha,C}^*$, we construct four integer grids in which `severity` and `age` span 5–45 and 55–90, respectively, one for each of the four combinations of levels of `sex` and `carrier`. We then simulate 100,000 draws from the joint posterior distribution (2.7) of the treatment–covariate interaction parameters, and use the 80th percentile as $\widehat{W}_{\alpha,C}^*$.

Figure 2.3 (left) displays the 80% credible subgroups with a threshold of $\delta = 0$. There is at least 80% posterior probability that the treatment effect is positive for all patients with covariate points in D , fully accounting for multiplicity and thus supporting regulatory approval for that subgroup. The observed covariate points are overlaid as the symbols \times (control arm) and $+$ (treatment arm). We see that we only have enough evidence to show that the oldest female patients with low-to-moderate severity benefit from the treatment versus control (at the 80% credible level). The PB and HPD methods yield similarly shaped exclusive credible subgroups that are larger and smaller, respectively, and the RCS method without the step-down procedure yields marginally smaller

Effect	Posterior Mean	Posterior SD	Sig.
(Intercept)	-2.45	1.72	
severity	0.64	1.03	
age	-2.18	1.36	
sex	4.04	2.35	
carrier	1.07	2.04	
sex:carrier	-4.60	3.29	
treatment	5.92	2.38	*
treatment:severity	-0.88	1.33	
treatment:age	3.49	1.61	*
treatment:sex	-4.28	2.66	
treatment:carrier	-1.50	2.46	
treatment:sex:carrier	-0.65	3.26	

Table 2.2: Posterior summaries of selected effect parameters. Continuous covariates are standardized. Estimates greater than 1.96 posterior standard deviations from 0 are marked significant.

subgroup than the one displayed. The uncertainty region $S \setminus D$ indicates characteristics of patients who may or may not benefit and for whom more evidence is needed. Patients in this region may be the focus of subsequent trials using enrichment designs [32]. A sensitivity analysis of a_0 and b_0 ranging from 1 to 10^5 resulted in nearly identical credible subgroups. Modifying \mathbf{R} to set prior variances for interaction parameters to a vague 100 also produced similar results, while shrinking interaction estimates even more strongly toward zero with prior variances of $1/100$ resulted in a larger exclusive credible subgroup. Additionally, placing a vague inverse-Wishart prior on \mathbf{R} centered at the value originally used gave results nearly identical to those obtained by using vague prior variances for interactions.

The right side of Figure 2.3 illustrates the results of a contrived analysis at the 50% credible level with benefit threshold $\delta = 2$ that includes, in addition to D and $S \setminus D$, the complement S^c of the inclusive credible subgroup. There is at least 50% posterior probability that the treatment effects for patients with covariate vectors in this region (here, younger male carriers with moderate-to-high severity) are simultaneously at most

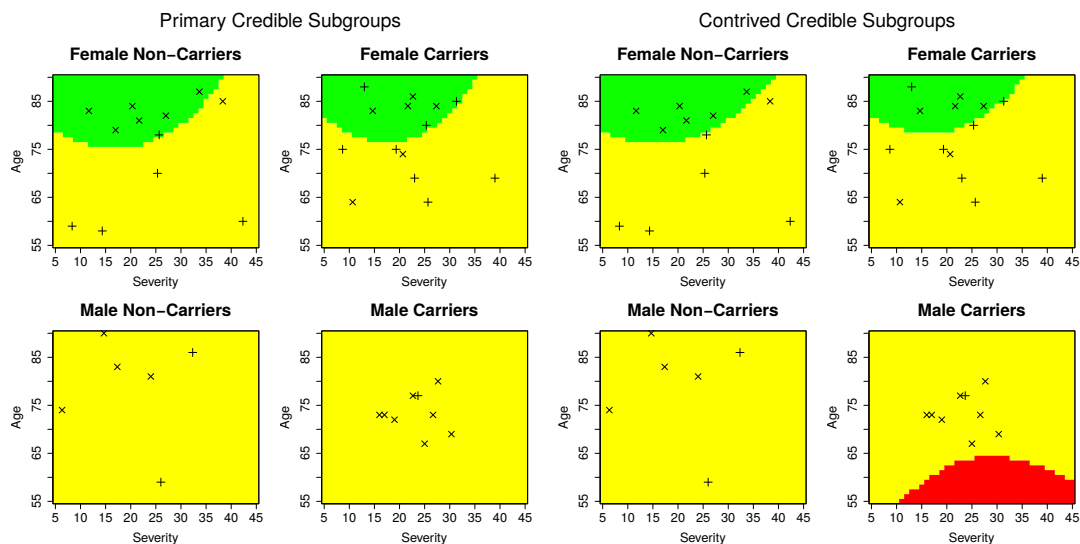


Figure 2.3: Step-down RCS credible subgroups at the 80% level with $\delta = 0$ (left), and at the 50% level with $\delta = 2$ (right). Patients in the control arm are represented by \times and those in the treatment arm by $+$.

δ , and investigators may consider abandoning efforts to show a beneficial treatment effect in this subgroup. However, note that S^G does not contain any data points, and is thus an extrapolation in this sense. Additionally, D contains only one patient in the treatment arm, and thus likely lacks sufficient common support between arms. These extrapolation issues arise primarily from the small sample size and rigidity of the linear model, and will be addressed in Chapters 3 and 5.

Figure 2.4 shows the contours of the posterior mean and standard deviation of the PTE surface. The linear contours of the mean surface and elliptical contours of the standard deviation surface combine to form the curved boundaries of the credible subgroups. Note that the mean surface continues to change at the same rate at the edges of the considered covariate space, while the standard deviation increases. Qualitatively, it is intuitive that the degree of certainty in any treatment effect decreases away from the primary mass of observed covariate points, though in the normal linear model the rate of change in the standard deviation may not be large enough to prevent the aforementioned problematic extrapolations.

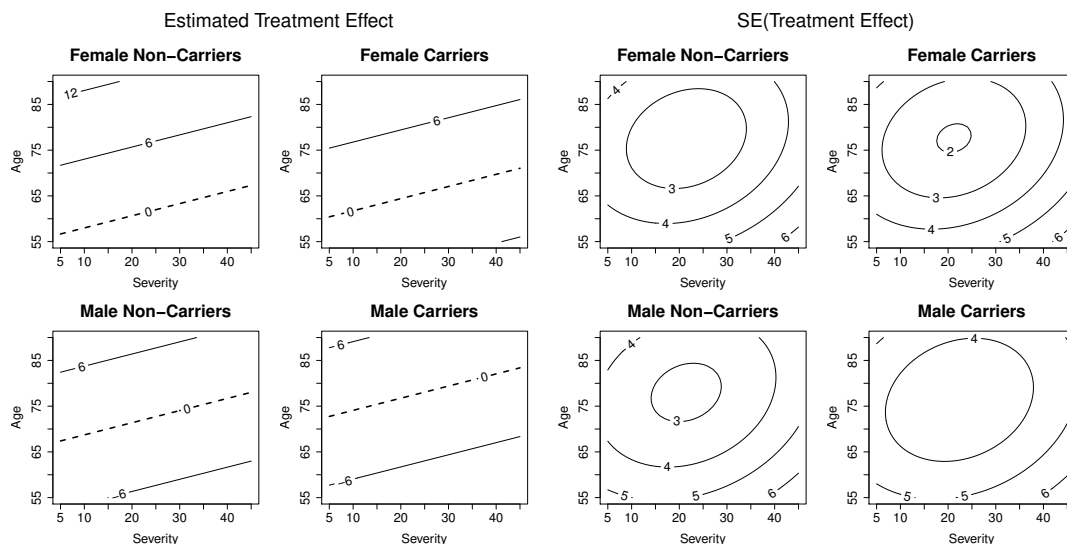


Figure 2.4: Contours of posterior mean and standard error surfaces, which combine to produce credible subgroups.

2.4 Discussion

The key advantage of the method of credible subgroups over existing methods for benefiting subgroup identification is the simultaneity of its conclusion: that there is high posterior probability that *all* members of the exclusive credible subgroup D have a personalized treatment effect exceeding δ , and *no* patients who are not members of the inclusive credible subgroup S have such a treatment effect. Such conclusions differ from those of the overall test: that the overall treatment effect exceeds δ , and, if the treatment effect is assumed to be homogeneous, it exceeds δ for everyone. The conclusions reached using credible subgroups are not necessarily more restrictive than those of the overall test: it may be the case that the overall treatment effect is not positive, but there is a substantial subgroup which benefits from treatment. Additionally, deferring classification of the uncertainty region until more evidence is obtained allows for stronger statements about the classifications already made.

Due to the two-step regression-classification procedure for determining credible subgroup pairs, the methods described in this chapter are extensible to non-normal and nonlinear models as long as it is possible to obtain a sample from the joint posterior of the predictive effect regression parameters or the personalized treatment effects directly,

though closed-form criteria for the HPD credible subgroups may not be available.

Another advantage of the credible subgroups approach is that it does not require prespecification of subgroups for testing, but only a list of covariates which may have predictive value. Additionally, the credible subgroups method more fully and naturally accounts for the dependence structure of the implicit hypothesis tests than do many methods of prespecified subgroups relying on Bonferroni or similar multiplicity adjustments, which are usually conservative in this context. However, credible subgroups are not as simple to describe as most prespecified subgroups, especially when there are multiple continuous predictive variables. Furthermore, the inclusion of a large number of predictive variables reduces power and makes interpretation and summarizing difficult.

The example analysis shows that although the sample sizes needed to detect benefiting populations are higher for credible subgroups methods than for analyses assuming homogeneous treatment effects, they are not as high as those typically needed for detecting heterogeneities as in traditional subgroup analysis. The example data of size $n = 41$ are sufficient to form a nonempty exclusive credible subgroup at the 80% level, but requires a level near 50% to identify effect heterogeneity in the form of the presence of both nonempty exclusive and nonuniversal inclusive credible subgroups. Significantly larger sample sizes may be necessary for meaningful results with confidence levels high enough to satisfy regulatory authorities. For example, in Section 3.3.2 a sample of size 369 yields a large exclusive credible subgroup at the 95% credible level, even under a more flexible semiparametric regression model.

Chapter 3

Credible Subgroups for General Regression Models

In the development of credible subgroups in Chapter 2, a normal linear model was used to provide some concrete expressions related to their construction. However, all three methods of construction generalize to arbitrary regression models as long as a sample from the joint posterior of all of the personalized treatment effects $\Delta(\mathbf{z})$ for \mathbf{z} in the covariate space C , denoted $\mathbf{\Delta}(C)$, can be obtained. Highest posterior density (HPD) credible subgroups may be constructed from a model parameterized by $\boldsymbol{\theta}$ by taking D and S to be the intersection and union, respectively, of all sets $B_{\boldsymbol{\theta}} \equiv \{\mathbf{z} : \Delta_{\boldsymbol{\theta}}(\mathbf{z}) > \delta\}$ for $\boldsymbol{\theta}$ in the HPD region of the appropriate credible level. However, the HPD region may need to be estimated from a Monte Carlo sample. Alternatively, the HPD region of the joint posterior of $\mathbf{\Delta}(C)$, may be used instead of that of $\boldsymbol{\theta}$. Such an approach is more computationally intensive, but allows nonparametric regression models to be used.

The restricted covariate space (RCS) and pure Bayes (PB) methods are based on simultaneous credible bands, for which we present generalized forms in Section 3.1. Next, Section 3.2 presents a general approach to estimating the benefiting subgroup using parametric models, including generalized linear models and models with variable selection, and Section 3.3 for semiparametric and nonparametric models. It will be seen that this variety of models can be used with identical machinery for constructing credible subgroups from a posterior sample of $\mathbf{\Delta}(C)$. Each of the discussions of GLMs, variable

selection, and semiparametric/nonparametric regression contain an example analysis using the RCS approach on one of our Alzheimer’s disease treatment trial datasets.

3.1 General forms of simultaneous credible bands

The simultaneous credible bands used to construct RCS credible subgroups (and in part for PB credible subgroups) for the normal linear model in Chapter 2 are given by

$$\mathbf{z}^\top \boldsymbol{\gamma} \in \mathbf{z}^\top \bar{\boldsymbol{\gamma}} \pm W_{\alpha, C}^* \sqrt{\mathbf{z}^\top \left(\frac{b}{a} \mathbf{H} \right) \mathbf{z}}, \quad (3.1)$$

where $W_{\alpha, C}^*$ is the $1 - \alpha$ quantile of the distribution of

$$W_C = \sup_{\mathbf{z} \in C} \frac{|\mathbf{z}^\top (\boldsymbol{\gamma} - \bar{\boldsymbol{\gamma}})|}{\sqrt{\mathbf{z}^\top \left(\frac{b}{a} \mathbf{H} \right) \mathbf{z}}}. \quad (3.2)$$

Since $\Delta(\mathbf{z}) = \mathbf{z}^\top \boldsymbol{\gamma}$, we can rewrite (3.1) and (3.2) as

$$\Delta(\mathbf{z}) \in \bar{\Delta}(\mathbf{z}) \pm W_{\alpha, C}^* \sqrt{\text{Var}[\Delta(\mathbf{z})]}, \quad (3.3)$$

and

$$W_C = \sup_{\mathbf{z} \in C} \frac{|\Delta(\mathbf{z}) - \bar{\Delta}(\mathbf{z})|}{\sqrt{\text{Var}[\Delta(\mathbf{z})]}}, \quad (3.4)$$

respectively. Intuitively, (3.3) defines the pre-image of $\{W_C \leq W_{\alpha, C}^*\}$, and therefore defines a $1 - \alpha$ simultaneous credible band for $\boldsymbol{\Delta}(C)$ produced by an arbitrary model. When the marginal posteriors of the $\Delta(\mathbf{z})$ belong to a location-scale family (as is often the case asymptotically), the particular form of the argument to the supremum in (3.4) standardizes the marginal posteriors to be identical but retains the dependence in the joint distribution, similarly to probabilistic copulae [33, 34].

In cases in which the marginal posteriors of the $\Delta(\mathbf{z})$ are not from a location-scale family, or location and scale otherwise poorly describe the distributions (e.g., discrete or highly asymmetric distributions), (3.3) is sub-optimal. An example of such a case will be considered in Chapter 4 with Bernoulli marginal posteriors. When (3.3) is not desirable, a quantile-based simultaneous credible band may be used. Let $F_\Theta(\theta) = \text{P}[\Theta \leq \theta]$ be the cumulative distribution function of θ , $G_\Theta(\theta) = \text{P}[\Theta < \theta]$ be its left-continuous

counterpart, and $F_\theta^{-1}(p) = \inf\{\theta : p \leq F_\Theta(\theta)\}$, $G_\Theta^{-1}(p) = \sup\{\theta : p \geq G_\Theta(\theta)\}$ be their inverses. We may then use the simultaneous credible band

$$\Delta(\mathbf{x}) \in \left[F_{\Delta(\mathbf{x})|\mathbf{y}}^{-1}(1 - W_{\alpha,C}^*), G_{\Delta(\mathbf{x})|\mathbf{y}}^{-1}(W_{\alpha,C}^*) \right], \quad (3.5)$$

with $W_{\alpha,C}^*$ set to be the $1 - \alpha$ quantile of the distribution of

$$W_C = \sup_{\mathbf{x} \in C} \max \{1 - F_{\Delta(\mathbf{x})|\mathbf{y}}[\Delta(\mathbf{x})], G_{\Delta(\mathbf{x})|\mathbf{y}}[\Delta(\mathbf{x})]\}. \quad (3.6)$$

The distribution and quantile functions of W_C may be estimated from the posterior sample. Detailed proofs of the correctness of (3.3) and (3.5) are presented in Appendix A.

3.2 Parametric regression

3.2.1 Generalized linear models

Generalized linear models (GLMs) are perhaps one of the simplest extensions of the normal linear model theory for credible subgroups. Although the conditional distribution of the responses is no longer necessarily normal, the underlying specification of the linear predictor can remain unchanged. This allows, for example, the model to retain the form

$$\eta = \mathbf{x}^\top \boldsymbol{\beta} + t\mathbf{z}^\top \boldsymbol{\gamma}, \quad (3.7)$$

with vague priors on main effects and conservative priors on interaction terms. Then $\Delta(\mathbf{z}) = \mathbf{z}^\top \boldsymbol{\gamma}$, and the general location-scale simultaneous credible band (3.3) may be used to construct credible subgroups from MCMC output or other methods of sampling the joint posterior of $\boldsymbol{\Delta}(C)$.

We turn to the adverse event data in the multi-endpoint dataset, concentrating on a comparison of Donepezil to the placebo. An analysis constructing credible subgroups according to what we propose as best practices (fixing a credible level a priori, using conservative priors for interaction parameters) yields an uncertainty region encompassing the entire covariate space, so we proceed with an example of the mechanism alone by using vague priors and computing maximum credible levels. Because the safety outcome is the presence or absence of at least one adverse event indicated by the physician to be possibly related to treatment, we use a logistic GLM. For patient i with prognostic and

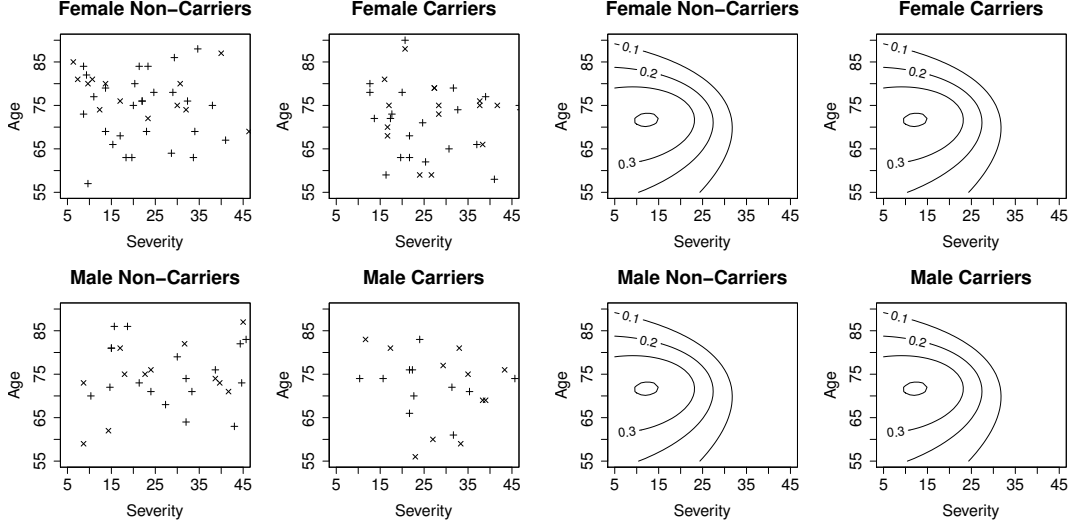


Figure 3.1: Observed covariate points (left, + for Donepezil, × for placebo) and maximum credible level contours for a non-inferiority (right, $\delta = -0.18$).

predictive covariate vectors \mathbf{x}_i and \mathbf{z}_i (including leading 1's) and treatment indicator $t_i = 1$ for the active control and 0 for the placebo control,

$$\begin{aligned}
 Y_i | \eta_i &\stackrel{iid}{\sim} \text{Bernoulli} [\text{logit}^{-1}(\eta_i)], \\
 \eta_i &= \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{\gamma}, \\
 \boldsymbol{\beta}_p, \boldsymbol{\gamma}_p &\stackrel{iid}{\sim} \text{Normal} [0, 10^4].
 \end{aligned}
 \tag{3.8}$$

If conservative priors were to be used, they would likely need to have smaller variances than those used for the normal linear model, as the inverse logit function approaches its supremum relatively quickly: $\text{logit}^{-1}(2.2)$ is already greater than 0.90. We fit the model using the NIMBLE [35] R package, retaining 100,000 draws after 10,000 burn-in iterations, and compute the maximum credible levels for inclusion in the exclusive credible subgroup for non-inferiority ($\delta = -0.18$, i.e., a 1.20 odds ratio).

Figure 3.1 (right) shows contours of the maximal credible level that permits assignment of covariate points to the exclusive credible subgroup. The figure indicates that the credible level must fall to near 40% in order for the exclusive credible subgroup for non-inferiority to be non-empty. Broadly speaking, the patients with low severity scores

are the first to be identified as not being unacceptably harmed, but the maximum credible levels are so low that it is questionable whether a non-empty “benefiting” subgroup exists at all.

3.2.2 Variable selection

Bayesian variable selection and model averaging may be incorporated into credible subgroups by using a regression model implementing such selection and either using the full posterior output (for model averaging), or selecting the most probable model and refitting the selected model (for strict selection). For example, stochastic search variable selection (SSVS) [36] specifies for each regression parameter a normal-normal mixture model with components chosen via a latent indicator variable. One high-variance component of the mixture represents a parameter that is included in the model, and the other, low-variance component represents a parameter being excluded. Other popular Bayesian variable selection methods include the least absolute shrinkage and selection operator (LASSO) [37] in its Bayesian form [38], an L1-penalized joint shrinkage estimator of the regression parameters.

As an example, we re-analyze the simple add-on therapy dataset from Chapter 2 via a modification of the SSVS model. For the likelihood and error variance, we use the same formulation as in Chapter 2, but define a spike-and-slab (or spike-and-bump) mixture prior for the treatment-covariate interactions:

$$\begin{aligned}
 \mathbf{Y} | \mathbf{X}, \mathbf{Z}, \mathbf{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2 &\sim \text{Normal} [\mathbf{X}\boldsymbol{\beta} + \mathbf{TZ}\boldsymbol{\gamma}, \sigma^2\boldsymbol{\Sigma}], \\
 \beta_p, \gamma_1 | \sigma^2 &\stackrel{iid}{\sim} \text{Normal} [0, 10^4 \sigma^2], \\
 \gamma_{p>1} | \sigma^2, \lambda_{p>1} &\stackrel{iid}{\sim} (1 - \lambda_p) \text{Dirac} [0] + \lambda_p \text{Normal} [0, 10\sigma^2], \\
 \lambda_p &\stackrel{iid}{\sim} \text{Bernoulli} [1/2], \\
 \sigma^2 &\sim \text{InverseGamma} [a_0, b_0],
 \end{aligned} \tag{3.9}$$

where the Dirac[0] distribution is the point mass at zero. This formulation differs slightly from the original SSVS formulation, which uses a very low-variance normal distribution instead of the Dirac function to maintain conjugacy, which is unnecessary here given the availability of modern MCMC tools. The conditional prior variance for the interaction terms is $10\sigma^2$ rather than the previously used σ^2 because the point mass

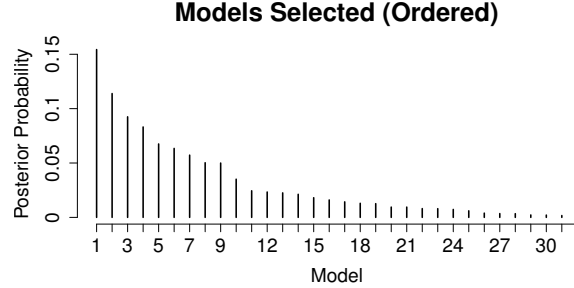


Figure 3.2: Credible subgroups using stochastic search variable selection and model averaging (left) and using the most frequently selected model only (right).

Posterior Probability	Predictive Covariates	Exclusive Credible Subgroup (80%)
0.15	age, sex	females 73–90
0.11	sex	females
0.09	age	ages 77–90
0.08	none	all
0.07	sex \times carrier status	females

Table 3.1: Top five models.

component replaces some of the shrinkage previously effected by the smaller variance. Sampling from the marginal posterior yields a Bayesian model-averaged result, while conditioning on the modal λ yields the result under the most frequently selected model. We use 100,000 MCMC samples after 10,000 burn-in iterations to produce the marginal model, and for each conditional model, 100,000 directly sampled draws from the posterior. Credible subgroups are produced at the 80% level using the RCS method with the step-down procedure. For the marginal model we use the quantile-based simultaneous credible bands (3.5) due to multi-modalities caused by the mixture priors.

Figure 3.2 shows the posterior probabilities of selected models in descending order. The most frequently selected model accounts for roughly a third more posterior probability than the runner-up, and the first five models (shown in Table 3.1) account for half of the posterior probability. The top five models tell a broadly consistent story that is also consistent with the all-variables credible subgroups produced in Section 2.3: the average treatment effect is driven by strong PTEs among older females.

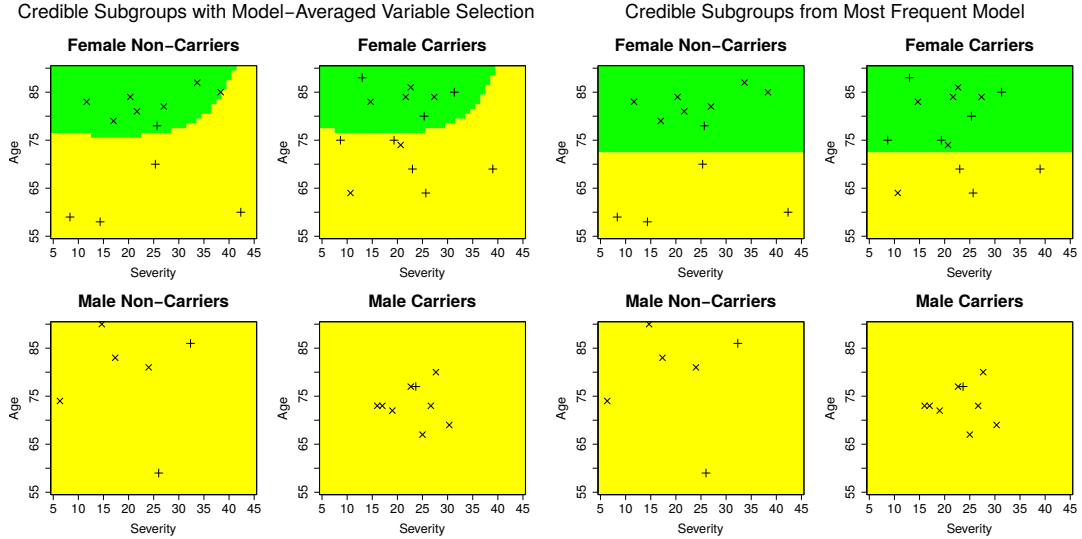


Figure 3.3: Credible subgroups (80%) using stochastic search variable selection and model averaging (left) and using the most frequently selected model only (right).

Figure 3.3 shows the model-averaged 80% credible subgroups from the marginal posterior (left) and from the most frequently selected model (using age- and sex-by-treatment interactions). The two plots tell the same broad story, though the conditional model yields a simpler and larger exclusive credible subgroup due to the smaller set of variables and resulting smaller variability in the PTE posterior. From a purely statistical standpoint, we would generally recommend using the marginal model, as it most faithfully represents the totality of uncertainty in the results. However, in this case the combination of consistency in the most probable models and much greater parsimony possible in reporting the credible subgroup results may warrant conditioning on the few most highly *a posteriori* probable models for “public” use.

3.3 Semiparametric and nonparametric regression

When making inferences about personalized treatment effects rather than average treatment effects, it is potentially much more important to allow sufficient flexibility in the regression model so as to ensure consistent estimation of the PTEs. To this end, several

nonparametric and semiparametric regression approaches for PTEs have been proposed, including random forests [39] in the virtual twins approach [16], Bayesian additive regression trees (BART) [12] in modeling for causal inference [40], and a hybrid approach defining linear treatment and baseline models using tree-based methods [41]. This section presents an approach for using such semiparametric and nonparametric models within the credible subgroups inferential framework, with special attention to penalized splines and BART. Section 3.3.1 provides a simulation study comparing the performance of linear, spline, and BART models, and Section 3.3.2 presents an example analysis using the multi-trial Alzheimer’s disease dataset.

This chapter focuses on normal-likelihood regression models, though the methods may be straightforwardly adapted to other likelihoods, such as GLM or proportional hazards survival models. Previously, a linear model of the form

$$E[Y|\mathbf{x}, \mathbf{z}; t] = \mathbf{x}^\top \boldsymbol{\beta} + t\mathbf{z}^\top \boldsymbol{\gamma}, \quad (3.10)$$

has been used, so that $\Delta(\mathbf{z}) = \mathbf{z}^\top \boldsymbol{\gamma}$. We first generalize to a semiparametric model based on the additive penalized spline model with factor-by-curve interactions [42]:

$$E[Y|\mathbf{x}, \mathbf{z}; t] = \beta_0 + \sum_{j=1}^p f_j(x_j) + t \left[\gamma_0 + \sum_{j=1}^p g_j(z_j) \right], \quad (3.11)$$

where the f_j and g_j are penalized cubic splines with radial bases and no intercepts:

$$f_j(x_j) = \beta_{j1}x_1 + \beta_{j2}x_j^2 + \beta_{j3}x_j^3 + \sum_{k=1}^{K_{fj}} u_{fjk} |x_j - \kappa_{fjk}|^3, \quad (3.12)$$

and the κ_{fjk} are fixed knots. The penalty is implemented by placing a Normal $[0, \sigma_{fj}^2]$ prior on the u_{fjk} and a vague InverseGamma $[10^{-4}, 10^{-4}]$ prior on the σ_{fj}^2 for conditional conjugacy. The g_j are specified similarly, using γ instead of β . We place flat priors on the β and γ , as specifying conservative priors on interaction terms that include knot random effect parameters is not as straightforward or interpretable as the corresponding priors in linear models. Models adding fixed effects or group-level random slopes and intercepts are straightforward to specify.

Model (3.11) can suffer from poorly identified parameters; however, the resulting $E[Y|\mathbf{x}, t]$, as well as quantities of the forms $E[Y|\mathbf{x}, \mathbf{z}; t = 1] - E[Y|\mathbf{x}, \mathbf{z}; t = 0]$ (the

PTE) or $E[Y|\mathbf{x}_1, \mathbf{z}_1; t] - E[Y|\mathbf{x}_0, \mathbf{z}_0; t]$, are typically stable. This stability, along with the tendency of software packages for other regression techniques to supply quantities of the form $E[Y|\mathbf{x}, \mathbf{z}; t]$, make $\Delta(\mathbf{z}) \equiv E[Y|\mathbf{x}, \mathbf{z}; t = 1] - E[Y|\mathbf{x}, \mathbf{z}; t = 0]$ a convenient definition of the PTE even when there cannot be an explicit separation of the treatment variable t from \mathbf{x} and \mathbf{z} in the model. For example, using the R package `BayesTree` to obtain a BART fit, we may concatenate the treatment indicator onto the covariate vector for each patient and fit $Y \sim \text{Normal}[\mu(\mathbf{w}), \sigma^2]$ where $\mathbf{w} = (x_1, \dots, x_p, z_1, \dots, z_p, t)$ with redundant entries removed. However, fully nonparametric models for which the PTE surface must be stored at every point in \mathcal{C} present challenges with respect to memory, as a sample from the posterior joint PTE distribution must be stored in an often very large (number of draws by number of covariate points) matrix.

3.3.1 Simulations

We perform a simulation study to evaluate certain frequentist properties of the credible subgroups generated by linear, spline, and BART models. A necessary property is valid (including conservative) coverage, i.e., $D \subseteq B \subseteq S$ at least $100(1 - \alpha)\%$ of the time. Given valid coverage, we compare regression models primarily by the sensitivity of D (i.e., how much of B is contained in D). We also evaluate the sensitivity of D under the step-down procedure relative to that under the single-step procedure.

We simulate 1000 data sets with $n = 100$ patients in each treatment arm. Results for simulations with $n = 25$, $n = 50$, and $n = 75$ are presented in Appendix B. Each subject i has covariate vectors $\mathbf{x}_i = \mathbf{z}_i = (1, z_{i2}, z_{i3})$ with $z_{i2} = 0, 1$ with equal probability and z_{i3} continuously uniformly distributed on $[-3, 3]$, a deterministic treatment assignment t_i , and a conditionally normally distributed response y_i . The covariates are used as both prognostic and predictive variables.

The outcomes are generated as $\text{Normal}[0 + \Delta(\mathbf{z}_i), 1]$ with $\Delta(\mathbf{z}_i)$ specified in the following six cases. In the null case, $\Delta(\mathbf{z}_i) = 0$. In the binary case, $\Delta(\mathbf{z}_i) = z_{i2}$. In the linear case, $\Delta(\mathbf{z}_i) = z_{i3}$. In the near-linear case, $\Delta(\mathbf{z}_i) = 2(\sqrt{z_{i3} + 3} - \sqrt{3})$. In the threshold case, $\Delta(\mathbf{z}_i) = \text{sign}(z_{i3})(z_{i3}/3)^{1/3} + 1/4$. In the non-monotone case, $\Delta(\mathbf{z}_i) = 1/2 - 3(z_{i3}/3)^2$. Table 3.2 includes graphical representations of these scenarios.

To each data set we fit a linear, spline, and BART model. For the linear and spline models, we place a vague `InverseGamma` $[10^{-4}, 10^{-4}]$ prior on the error variance and







Data Generating Mechanism	Diagram of $\Delta(\mathbf{x})$	Model	Coverage	Sensitivity of D	Step-Down Efficiency
Null Effect		Linear	0.89	—	—
		Spline	0.92	—	—
		BART	0.99	—	—
Binary		Linear	0.91	0.97	1.01
		Spline	0.95	0.56	1.05
		BART	0.98	0.82	1.05
Linear		Linear	0.88	0.90	1.02
		Spline	0.94	0.77	1.09
		BART	1.00	0.70	1.08
Near-Linear		Linear	0.75	0.71	1.04
		Spline	0.97	0.39	1.19
		BART	1.00	0.27	1.25
Threshold		Linear	0.61	0.85	1.03
		Spline	0.96	0.52	1.07
		BART	0.99	0.48	1.07
Non-Monotone		Linear	0.23	0.52	1.10
		Spline	0.96	0.63	1.05
		BART	0.97	0.46	1.08

Table 3.2: Simulation study results. Operating characteristics of 80% credible subgroups with $n = 100$ patients in each study arm. Struck-through sensitivities indicate insufficient coverage and should be treated with caution.

flat priors on fixed effect coefficients. For the spline model we place InverseGamma [2, 1] shrinkage priors on the random effect variances. The BART model is fit using the default settings in the R package `BayesTree`. All Gibbs samplers are run for 100 burn-in and 1000 retained iterations, which appears to be acceptable for these simple models.

To determine credible subgroups at the 80% credible level, we use as the covariate space the grid in which $z_1 = 1$, $z_2 = 0, 1$, and z_3 ranges from -3 to 3 in increments of 0.1 . We compute the result under both the single-step and step-down procedures using the location-scale simultaneous credible band (3.3).

Table 3.2 displays the average summary statistics for 80% credible subgroups for each model and data generating mechanism (DGM) at $n = 100$ patients per arm. Each

model has sufficient coverage, except for the linear model in the non-linear cases. Given sufficient coverage, the sensitivity of D will usually be the driving factor in choosing a model. In this regard, the spline model performs better than BART in all cases except when the binary covariate drives the treatment effect heterogeneity, but the spline model’s advantage is small in the “threshold” case in which the continuous covariate behaves similarly to a binary one. When the true treatment effect heterogeneity is linear, the linear model outperforms both with respect to sensitivity of D. Generally, we recommend the spline model when continuous predictive covariates are present, as even a small departure from linearity can render the coverage of the linear model insufficient (see the “near-linear” case). Finally, the step-down method consistently improved the sensitivity of D, sometimes to a large extent (nonparametric fits in the near-linear case).

In the absence of model mis-specification, all methods appear to have conservative coverage. In fact, the realized error rate does not rise far past $\alpha/2$. The relevance of $\alpha/2$ as an error rate here is that credible subgroups can only make an error in one direction at each covariate point: if the true PTE is positive, the only error is under-estimation, while if the true PTE is non-positive, the only error is over-estimation; however, we cannot know a priori the sign of the PTE. Thus the apparent conservatism is due to the fact that none of the displayed simulations represent the worst-case scenario that the procedure protects against: a small-magnitude treatment effect that crosses zero frequently.

3.3.2 Semiparametric example analysis

We consider data from a sequence of four clinical trials for Alzheimer’s disease (AD) treatments carried out by AbbVie, all of which include arms for a placebo and the same “standard of care” treatment. We compare the standard of care to the placebo with respect to change in disease severity over 12 weeks, using data from all four trials. Combined, the studies are comprised of 369 complete-case patients from 9 countries.

We consider six baseline patient characteristics: disease severity, change in disease severity during run-in, long-term cognitive decline rate, age, carrier status of the ApoE4 allele, and sex. The change in severity score in the 3–4 week run-in period between screening and randomization, which we call **prechange**, is included as a main effect in an attempt to adjust for the “learning effect” in which patients become familiar

with the ADAS-Cog 11 instrument; however it is not included as a predictive covariate because it is not thought to be useful for practitioners due to its high variability and delaying of treatment. We do include as a predictive covariate the long-term cognitive decline rate, `drate`, which is defined as the total drop in score on the Mini Mental State Examination (MMSE) divided by the time, in years, since onset of first symptoms. Since `age` was highly correlated with `drate`, and the inclusion of `age` in the model for the present analysis was detrimental to penalized model fit as evaluated by DIC and LPML, we excluded it from our analysis. The outcome, `improvement`, is the negative change in severity (baseline minus end-of-study), so that a positive value represents a good outcome.

Our outcome model may be broadly summarized (in R-like syntax) as

$$\begin{aligned}
 \text{improvement} \sim & \text{Intercept} + \text{r(country)} + \text{sex} + \text{carrier} \\
 & + \text{f(prechange)} + \text{f(severity)} + \text{f(drate)} \\
 & + \text{treatment} + \text{treatment:r(country)} \\
 & + \text{treatment:sex} + \text{treatment:carrier} \\
 & + \text{treatment:f(severity)} + \text{treatment:f(drate)}
 \end{aligned} \tag{3.13}$$

where `r(·)` represents a traditional centered, normally-distributed random intercept or slope, `f(·)` represents a penalized cubic spline, colons denote interactions, and errors are normally distributed. We place knots for spline terms at increments of 2 across the observed ranges of `prechange` and `severity`, and at increments of 1 for `drate`. Variances for error, random effects, and penalized spline coefficients are given vague `InverseGamma(0.001, 0.001)` priors. Fixed effects are given flat priors. We report results only in a subset of the covariate space which has sufficient observation density and common support, discussed further in Section 5.1.2. Restricting `severity` to `[9, 49]` and `drate` to `[0, 8]` we include this entire region and exclude less than 9% of patients, while reducing the size of the covariate space by more than half.

The model was fit by Gibbs sampling using 100,000 iterations after 1000 burn-in iterations. Convergence appears near-immediate and mixing good from trace plots, and the time series-based effective sample sizes for most coefficients are above 75,000. Effective sample size was lowest for some country random effects with few patients, and for certain random effect variances.

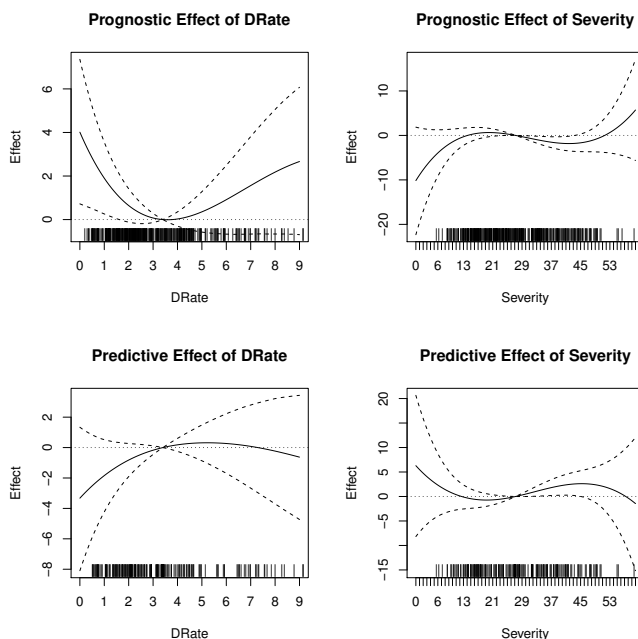


Figure 3.4: Estimated nonlinear effects and 95% pointwise credible bands on standardized covariate scale, relative to sample mean. Rug plot represents observed covariate distribution.

Figure 3.4 displays the nonparametrically fitted effect curves. The posterior mean effects display substantial nonlinearity and even nonmonotonicity, though only the credible band for the prognostic effect of severity gives convincing evidence of nonlinearity at the 95% credible level, and none are convincingly nonmonotone by the same criterion.

Figure 3.5 displays the credible subgroups at the 95% level, using the step-down testing procedure (Algorithm 2). The exclusive credible subgroup generally contains patients with high severity and rate of decline. The exclusive credible subgroup in this case includes approximately 17% more cells than the corresponding exclusive credible subgroup when the full observed ranges of severity and d-rate are used, and approximately 3% more than when the single-step procedure is used.

We also fit two other models: a version of (3.13) in which the penalized spline terms were replaced with linear effects, and the default implementation of BART from the R package `BayesTree` for 100,000 iterations, thinned to 10,000 iterations due to memory considerations, after 1000 burn-in iterations. Figure 3.6 displays the 95% credible

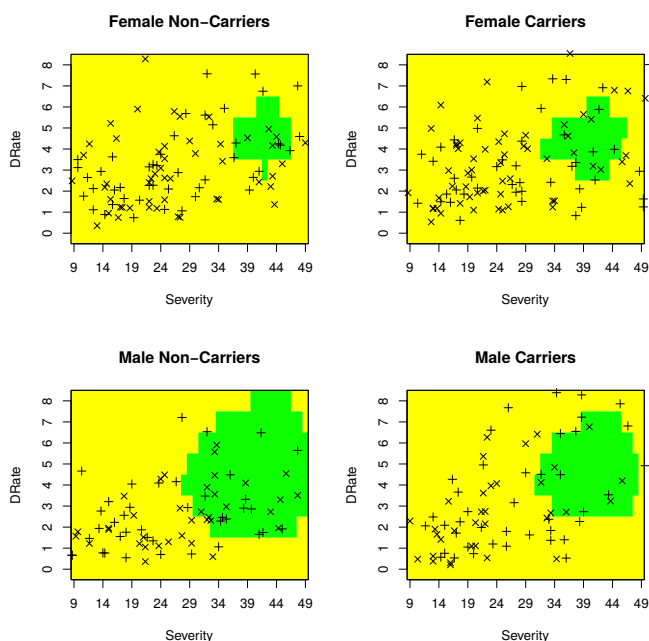


Figure 3.5: Credible subgroups at the 95% level. Green points represent the exclusive credible subgroup, and yellow the remainder of the inclusive credible subgroup.

subgroups for the linear and BART models, and Figure 3.7 compares the posterior mean PTE surface between (3.13) and its linear counterpart. Such visualizations of the estimated PTE surface may be useful to trialists who wish to more fully understand possible nonlinear features of the surface that would be lost under a linear model. As may be expected, the linear model simplifies the estimated PTE surface, which, along with the variances of the PTEs, yield a smoother exclusive credible subgroup. By contrast, BART, which divides the covariate space into rectangular cells having constant PTE in each cell, yields a more rectangular exclusive credible subgroup. Due to the superior performance of the spline-based model in the simulation study for the nonlinear continuous effect case, we promote those Figure 3.5 credible subgroups (which are intermediate to those in Figure 3.6) as the best choice. The resulting credible subgroups are consistent with the observation that the linear model fits poorly while the BART fit is too conservative in similar situations.

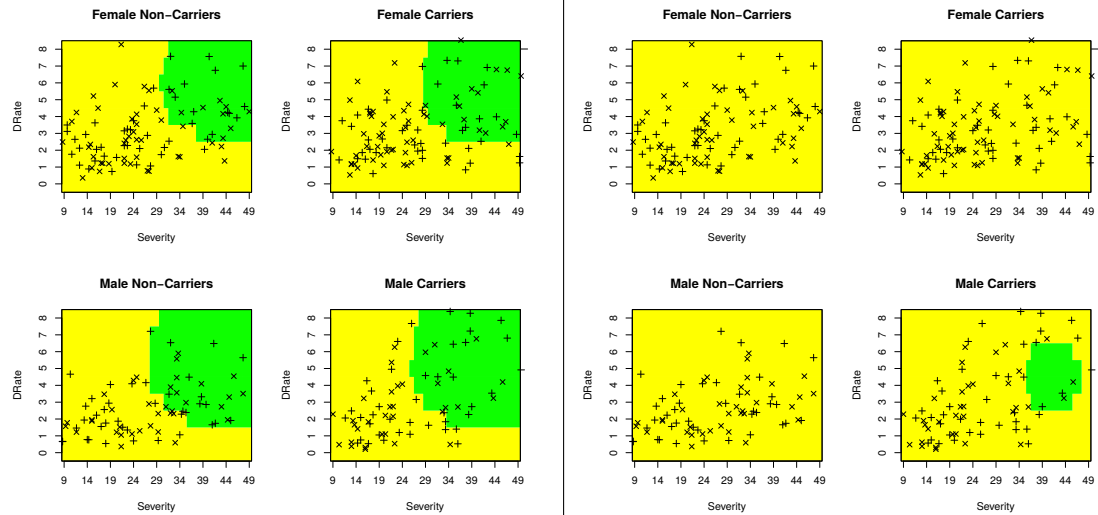


Figure 3.6: Credible subgroups at the 95% level for the linear (left) and BART (right) models.

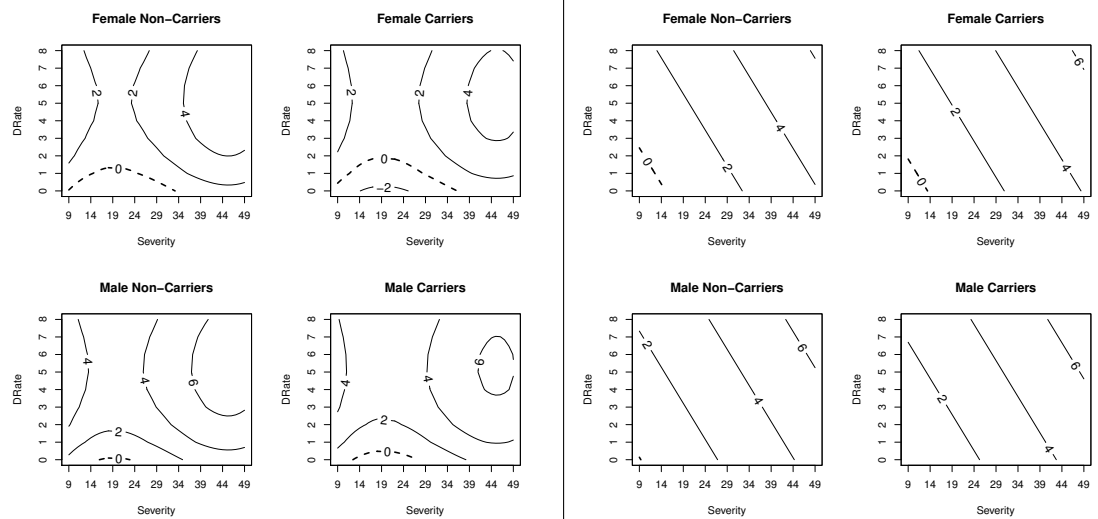


Figure 3.7: Posterior mean PTE surfaces for the semiparametric (left) and linear (right) models.

3.4 Discussion

Because of the relative independence of the credible subgroups inferential tools and the selected regression method—all that is needed is a sample from the joint posterior of the personalized treatment effects—the choice of parametric, semiparametric, or nonparametric methods may be made with full focus on flexibility and applicability to the problem, rather than being muddled by technical considerations about the associated inferential process. Such freedom may make “black box” nonparametric methods such as BART appealing for their flexibility, but also allows the use of more interpretable models such as additive spline models if desired. However, increased flexibility of semiparametric and nonparametric regression models come at a cost in terms of power, due to the looser dependencies of the PTEs across the covariate space.

Variable selection techniques do, however, raise some important inferential questions. Within the Bayesian inferential framework the analysis often naturally yields a marginal posterior or model-averaged result, incorporating information from many models at once rather than conditioning on a most probable model. In a “purist” sense, the marginal posterior and the credible subgroups produced from them constitute the most faithful representation of uncertainty in the analysis. Such exact results may be desired for use in a computer program, but conditional models incorporating only a few variables may be more practical when attempting to provide parsimonious recommendations to non-statisticians or even the general public.

Chapter 4

Subgroup Inference with Multiple Endpoints and Many Treatments

Previous chapters have dealt with benefiting subgroup identification under the assumption that it is straightforward to define what “benefit” actually means. This is generally the case when one test treatment is being compared to one control with respect to a single endpoint. However, many treatments affect more than one facet of patient well-being, and it is not always possible or even desirable to fix one definition of benefit for all patients. Additionally, treatment developers may be interested in testing the treatment effect with respect to multiple endpoints in a way that would pass regulatory review. Finally, some studies aim to evaluate multiple test treatments or doses of a single test treatment, and multiple control arms may be used when the standard of care is not well-established or non-inferiority is being tested.

This chapter develops subgroup analysis methods to handle cases in which more than two treatments are being compared with respect to multiple endpoints. This multivariate problem setting admits several ways of defining a treatment effect and benefiting subgroup, as well as strategies for choosing the multiplicities for which to adjust. The initial discussion is general and can be applied with any method that reports (or can potentially yield) an estimate for a benefiting subgroup B , especially when inference includes posterior probabilities $P[D \subseteq B | \mathbf{y}]$. Here D is the reported estimate, together with the posterior probability that the reported subset does indeed characterize

covariate combinations with a substantially higher treatment effect (or one exceeding some other threshold). The discussion can also apply to methods which produce some subgroup by any means and then tests for a within-subgroup treatment effect, though these methods will not be our focus here. Eventually, in the implementation we will incorporate the credible subgroups approach.

A related course of research is underway in the area of dynamic treatment regimes (DTRs), which infers optimal processes in which sequences of treatments are given to a single patient in a response-adaptive manner. Several methods have been developed to select the best from among many previously vetted treatments for individual patients in the presence of multiple relevant endpoints. Response, non-response, and death may be treated as levels of an ordinal outcome, with the trade-off between response and death quantified using a real-valued utility function elicited from experts [43]. Additionally, a trial in which four treatments were tested in a two-stage regime has been reported [44]. More general treatments of multi-endpoint approaches have been considered, including using patient preferences among various endpoints in addition to clinical characteristics in the estimated treatment rule [45], methods for identifying optimal treatment regimes for all linear combinations of endpoints [46], and reporting treatment regimes with sets of non-inferior treatment choices [47, 48]. Since research in DTRs focuses on providing optimal care to a given patient, attention is not generally paid to Type I error control. In contrast, our work focuses on single-stage, population-level inferences for a given treatment, and owing to our focus on the regulatory process, attention must be paid to Type I error and its control under multiplicity of endpoints, treatments, and covariate profiles.

The remainder of the chapter is organized as follows. Section 4.1 develops an inferential framework for trials with more than two arms and multiple endpoints, with Section 4.1.3 extending the concept of credible subgroups in this setting. Section 4.2 evaluates the proposed methods with respect to sensitivity, specificity, and Type I error via a simulation study, while Section 4.3 illustrates the use of a subset of the methods on the multi-endpoint trial dataset. Finally, Section 4.4 offers closing remarks.

4.1 Subgroup inference

4.1.1 Multiple endpoints

Results regarding the effect of a treatment on a specific endpoint are generally not considered in a vacuum. For example, an experimental treatment may have approximately the same effect as the standard of care on the primary endpoint (cognitive function score in our example), but have a lower instance of adverse side effects such as nausea. In such a situation, it would be useful to know not only who benefits from the experimental treatment with respect to the primary endpoint, but also who is likely to avoid side effects.

Suppose that there are $K \geq 2$ endpoints by which the test treatment is being compared to the control, and let $\Delta_k(\mathbf{z})$ be the treatment effect at covariate point \mathbf{z} with respect to the k th endpoint. It is possible to construct subgroup inferences for the treatment effect corresponding to each endpoint, either independently or adjusting for the multiplicity of endpoint inferences. For a set of independently estimated subgroups $\{D_k\}_{k=1}^K$, we have that for each endpoint k and covariate point $\mathbf{z} \in D$, $P[\mathbf{z} \in B_k | \text{data}] \geq 1 - \alpha$. A set of subgroups is *simultaneous* (adjusting for endpoint multiplicity) if $P[\{k : \mathbf{z} \in D_k\} \subseteq \{k : \mathbf{z} \in B_k\} | \mathbf{y}] \geq 1 - \alpha$ for each \mathbf{z} . Both methods result in K subgroup estimates, and may be used when each of the endpoints is of interest separately, rather than in combination.

A way to construct a single subgroup estimate that incorporates information about each of the endpoint effects is through a *utility function*, e.g., trading off probability of response and risk of death. Let u be some utility function of all the endpoints, and define the treatment effect $\Delta_u(\mathbf{z})$ as $E[u | \mathbf{z}, t = 1] - E[u | \mathbf{z}, t = 0]$, where $t = 1$ indicates the test treatment and $t = 0$ the control. The benefiting subset B and the subgroup estimate D may then be defined in the same way as in the single-endpoint case. Constructing a single subgroup estimate may simplify interpretation, but it is often difficult for multiple parties to agree on a single, often stylized utility function, especially for diseases such as Alzheimer's that affect quality of life in complex ways and drugs that frequently have uncomfortable side effects. If a range or distribution U of utility functions is to be considered, $\Delta_U(\mathbf{z})$ may be constructed to reflect some summary of the distribution of the $\Delta_u(\mathbf{x})$ as u varies, such as the mean, median, or minimum.

We can also construct a joint subgroup report motivated by the decision-theoretic concept of *admissibility*. Recall that a decision rule is admissible if there are no other rules that always perform at least as well and better in at least one case. Here we would like to call a test treatment *admissible at \mathbf{z}* if the control treatment does not perform at least as well with respect to *every* endpoint and better with respect to at least one endpoint for a patient with covariate vector \mathbf{z} . Strictly speaking, the formalization of this definition is that a treatment is admissible at \mathbf{z} unless $\Delta_k(\mathbf{z}) \leq 0$ for all k and the inequality is strict for at least one k .

Next, we generalize to allow for thresholds of clinical significance and noninferiority. In addition to the δ_k , the thresholds for clinical significance, let $\epsilon_k \leq \delta_k$ be thresholds for non-inferiority, i.e., a treatment is considered “just as good” if $\epsilon_k \leq \Delta_k(\mathbf{z}) \leq \delta_k$. Introducing these thresholds allows for multiple formulations of criteria. We call a treatment *weakly admissible at \mathbf{z}* if $\Delta_k(\mathbf{z}) > \delta_k$ for at least one k or $\Delta_k(\mathbf{z}) \geq \epsilon_k$ for all k . This is the generalization of our previous definition of admissibility most directly related to the decision-theoretic concept. However, a treatment may be undesirable if it is demonstrably inferior with respect to one endpoint, even if it is superior in others, or if it is not superior in any. Thus we call a treatment *strongly admissible at \mathbf{z}* if $\Delta_k(\mathbf{z}) > \delta_k$ for at least one k and $\Delta_k(\mathbf{z}) \geq \epsilon_k$ for all k . A related method is to require only *noninferiority at \mathbf{z}* , i.e., that $\Delta_k(\mathbf{z}) \geq \epsilon_k$ for all k .

The decision-theoretic criteria described above may be written in notation unified with the previous formulations of individual-endpoint and utility function treatment effects. For example, if we define $\mathbb{I}(\text{condition})$ to be 1 if the condition is true and 0 otherwise, an indicator of strong admissibility may be written as

$$\Delta_{sa}(\mathbf{z}) = \mathbb{I} \left[\max_k \{ \Delta_k(\mathbf{z}) - \delta_k \} > 0 \right] \mathbb{I} \left[\min_k \{ \Delta_k(\mathbf{z}) - \epsilon_k \} \geq 0 \right] \quad (4.1)$$

and compared to $\delta_{sa} = 0$ (with *sa* indicating strong admissibility) in the same fashion as the treatment effects above. We may similarly define the indicator $\Delta_{wa}(\mathbf{z})$ for weak admissibility (*wa*), which would then be compared to $\delta_{wa} = 0$. We can then define B as the set of \mathbf{z} for which the treatment is admissible, and construct the desired joint subgroup report in the usual fashion. We term this approach the *direct method* for estimating admissibility.

A multiplicity problem arises when constructing subgroup reports from Δ_{sa} or Δ_{wa} .

As more endpoints are included in an analysis, the frequentist probability of identifying at least one endpoint with respect to which the test treatment is superior or inferior increases, even when treatments are equivalent with respect to every endpoint. This makes it more likely for a treatment to be classified as weakly admissible or not strongly admissible. To avoid these biases, we may construct admissibility inferences via a *fully adjusted method* as follows. Let $\{D_k\}_{k=1}^K$ be a simultaneous set of subgroup reports for superiority with respect to the K endpoints such that for all \mathbf{z} , $P(\{k : \mathbf{z} \in D_k\} \subseteq \{k : \mathbf{z} \in B_k\} | \text{data}) \geq 1 - \alpha$, and $\{D'_k\}_{k=1}^K$ be similarly defined for non-inferiority. Then for weak and strong admissibility, respectively,

$$D_{wa} = \left\{ \bigcup_{k=1}^K D_k \right\} \cup \left\{ \bigcap_{k=1}^K D'_k \right\}, \quad \text{and} \quad D_{sa} = \left\{ \bigcup_{k=1}^K D_k \right\} \cap \left\{ \bigcap_{k=1}^K D'_k \right\}. \quad (4.2)$$

4.1.2 Multiple endpoints and many treatments

Suppose now that there are $M > 2$ treatments being considered. It may not be desired to compare every treatment to every other. For example, we may envision a scenario in which there are three test treatments and one control, and it is desired to determine for each test treatment which patients benefit relative to the control. Consider a *competition graph* $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{t = 1, \dots, M\}$ is the set of treatment arm vertices and $\mathcal{E} = \{(t, c)\}$ is the set of directed edges where (t, c) is present if treatment t is being compared to control c . Let $\mathcal{E}(t)$ be the set of edges which originate at t . Let $\Delta_k^{tc}(\mathbf{z})$ be the effect of treatment t relative to treatment c for endpoint k , and δ_k^{tc} be a threshold of clinical significance such that $\Delta_k^{ct}(\mathbf{z}) = -\Delta_k^{tc}(\mathbf{z})$ but δ_k^{ct} is not necessarily the same as δ_k^{tc} . We generalize each of the two-arm methods to the many-arm multiple-endpoint case.

A subgroup inference may be constructed for each of the $K|\mathcal{E}|$ endpoint-comparison combinations, either independently or simultaneously (adjusting for multiplicity among endpoints and comparisons). For a set of independently generated inferences $\{D_k^{tc}\}$, we have that for each endpoint-comparison pair $(k, (t, c))$ and covariate point $\mathbf{z} \in D_k^{tc}$, $P[\mathbf{z} \in B_k^{tc} | \mathbf{y}] \geq 1 - \alpha$. For a simultaneous set of inferences we require that for each \mathbf{z} , $P[\{(k, (t, c)) : \mathbf{z} \in D_k^{tc}\} \subseteq \{(k, (t, c)) : \mathbf{z} \in B_k^{tc}\} | \mathbf{y}] \geq 1 - \alpha$, where (t, c) varies over \mathcal{E} . These methods may be useful when each of the endpoints and treatment-comparisons are of interest separately.

Alternatively, inferences may be constructed for each of the KM endpoint-treatment

combinations, in which each treatment t is compared against the totality of its competition, the comparison being denoted as t^* . Again, the estimates $D_k^{t^*}$ may be determined independently or simultaneously. Let $\Delta_k^{t^*}(\mathbf{x}) = \min_{c \in \mathcal{E}(t)} \{\Delta_k^{tc}(\mathbf{z}) - \delta_k^{tc}\}$ be the treatment effect versus the totality of competition and $\delta_k^{t^*} = 0$ be the corresponding threshold, so that t is considered beneficial if it outperforms all of its competition by the corresponding margins. For independent sets of pairs, we require that for each (k, t) and $\mathbf{z} \in D_k^{t^*}$, $P[\mathbf{z} \in B_k^{t^*} | \mathbf{y}] \geq 1 - \alpha$. For a simultaneous set of inferences, we would require for each \mathbf{z} , $P[\{(k, t) : \mathbf{z} \in D_k^{t^*}\} \subseteq \{(k, t) : \mathbf{z} \in B_k^{t^*}\} | \mathbf{y}] \geq 1 - \alpha$.

Utility functions may be used to reduce the effective number of endpoints to one, and either $|\mathcal{E}|$ inferences may be constructed for pairwise treatment effects Δ_u^{tc} , or M may be constructed for the treatment effects $\Delta_u^{t^*}$. Alternatively, inferences for weak and strong admissibility or noninferiority may be constructed, either for a treatment against each of its competitors separately (e.g. with respect to each Δ_{sa}^{tc}), or for a treatment against the totality of its competition (e.g. with respect to $\Delta_{sa}^{t^*}$). Again, sets of credible subgroup pairs may be constructed independently or simultaneously. If using admissibility inferences corrected for multiplicity as in (4.2), a similar multiplicity adjustment may be made for many arms by taking, for weak and strong admissibility, respectively,

$$D_{wa}^{t^*} = \bigcap_{(t,c) \in \mathcal{E}(t)} D_{wa}^{tc}, \quad \text{and} \quad D_{sa}^{t^*} = \bigcap_{(t,c) \in \mathcal{E}(t)} D_{sa}^{tc}. \quad (4.3)$$

4.1.3 Credible subgroups

We now develop in detail the implementation of the general approach for the adjustment for multiple endpoints and multiple treatment comparisons when the underlying model is the report of credible subgroup pairs as proposed previous chapters. This implementation is particularly interesting because it simplifies the form of certain probability statements by adjusting for multiplicity not only of endpoints and treatments, but covariate points as well.

Recall that an exclusive credible subgroup D and an inclusive credible subgroup S constitute a credible subgroup pair (D, S) if the posterior probability that $D \subseteq B \subseteq S$ is at least $1 - \alpha$, i.e. $P[D \subseteq B \subseteq S | \mathbf{y}] \geq 1 - \alpha$. When considering multiple endpoints and many treatments, the appropriate probability statements satisfied by the construction

of credible subgroups are $P [D_k^{tc} \subseteq B_k^{tc} \subseteq S_k^{tc} | \mathbf{y}] \geq 1 - \alpha$ for independent pairs, and $P [\forall (k, (t, c)) \in \{1, \dots, K\} \times \mathcal{E}, D_k^{tc} \subseteq B_k^{tc} \subseteq S_k^{tc} | \mathbf{y}] \geq 1 - \alpha$ for simultaneous pair sets.

A simultaneous set of credible subgroup pairs is derived from the joint distribution of many treatment effects corresponding to various covariate points, endpoints, and treatment comparisons. Let $\bar{\Delta}_k^{tc}(\mathbf{z}) = E [\Delta_k^{tc}(\mathbf{z}) | \mathbf{y}]$. Simultaneous credible bands for the $\Delta_k^{tc}(\mathbf{z})$ on C may be constructed as

$$\Delta_k^{tc}(\mathbf{z}) \in \bar{\Delta}_k^{tc}(\mathbf{z}) \pm W_{\alpha, C}^* \sqrt{\text{Var}[\Delta_k^{tc}(\mathbf{z})]} \quad (4.4)$$

where $W_{\alpha, C}^*$ is the $1 - \alpha$ quantile of the distribution of

$$W = \sup_{(\mathbf{z}, k, (t, c))} \frac{|\Delta_k^{tc}(\mathbf{z}) - \bar{\Delta}_k^{tc}(\mathbf{z})|}{\text{Var}[\Delta_k^{tc}(\mathbf{z})]}. \quad (4.5)$$

and $(\mathbf{z}, k, (t, c))$ ranges over $C \times \{1, \dots, K\} \times \mathcal{V}$. The value of W_{α}^* may be estimated from a sample from the joint posterior of the $\Delta_k^{tc}(\mathbf{z})$.

The use of (4.4) is most appropriate when the posterior distributions of the $\Delta_k^{tc}(\mathbf{z})$ are continuous and differ only by a scale parameter. When discontinuous posterior distributions are present, for instance that of $\Delta_{sa}(\mathbf{z})$ in (4.1), a quantile-based credible band may be more appropriate. Let $F(y) = P[Y \leq y]$, $F^{-1}(p) = \inf \{y : p \leq F(y)\}$, $G(y) = P[Y < y]$, and $G^{-1}(p) = \sup \{y : p \geq G(y)\}$. If W_{α}^* is the $1 - \alpha$ quantile of the distribution of

$$W = \sup_{(\mathbf{z}, k, (t, c))} \min \left\{ 1 - F_{\Delta_k^{tc}(\mathbf{z})} [\Delta_k^{tc}(\mathbf{z})], G_{\Delta_k^{tc}(\mathbf{z})} [\Delta_k^{tc}(\mathbf{z})] \right\}, \quad (4.6)$$

then

$$\Delta_k^{tc}(\mathbf{z}) \in \left[F_{\Delta_k^{tc}(\mathbf{z})}^{-1}(1 - W_{\alpha}^*), G_{\Delta_k^{tc}(\mathbf{z})}^{-1}(W_{\alpha}^*) \right] \quad (4.7)$$

is a $1 - \alpha$ simultaneous credible band. Distribution functions and W_{α}^* may be estimated from a sample from the joint posterior of the $\Delta_k^{tc}(\mathbf{z})$.

Given simultaneous credible bands such as those in (4.4) and (4.7), the exclusive credible subgroups D_k^{tc} and inclusive credible subgroups S_k^{tc} are constructed by comparing the upper and lower bounds of the bands to δ_k^{tc} . In the case of (4.4), the exclusive credible subgroup D_k^{tc} and inclusive credible subgroup S_k^{tc} are given by

$$\begin{aligned} D_k^{tc} &= \left\{ \mathbf{z} \in C : \bar{\Delta}_k^{tc}(\mathbf{z}) - W_{\alpha, C}^* \sqrt{\text{Var}[\Delta_k^{tc}(\mathbf{z})]} > \delta_k^{tc} \right\}, \\ S_k^{tc} &= \left\{ \mathbf{z} \in C : \bar{\Delta}_k^{tc}(\mathbf{z}) + W_{\alpha, C}^* \sqrt{\text{Var}[\Delta_k^{tc}(\mathbf{z})]} \geq \delta_k^{tc} \right\}, \end{aligned} \quad (4.8)$$

and $P[D_k^{tc} \subseteq B_k^{tc} \subseteq S_k^{tc} | \mathbf{y}] \geq 1 - \alpha$. The loose inequality is used for S_k^{tc} so that if $\delta_k^{tc} = 0 = \delta_k^{ct}$ then $D_k^{ct} = (S_k^{tc})^c$, the complement of the opposite comparison's inclusive subgroup. Credible subgroups derived from the form (4.7) are constructed similarly.

Once the (D_k^{tc}, S_k^{tc}) are available, credible subgroups for admissibility may be constructed through the following analogs of equations (4.2) and (4.3):

$$\begin{aligned} (D_{wa}, S_{wa}) &= \left(\left\{ \bigcup_{k=1}^K D_k \right\} \cup \left\{ \bigcap_{k=1}^K D'_k \right\}, \left\{ \bigcup_{k=1}^K S_k \right\} \cup \left\{ \bigcap_{k=1}^K S'_k \right\} \right), \\ (D_{sa}, S_{sa}) &= \left(\left\{ \bigcup_{k=1}^K D_k \right\} \cap \left\{ \bigcap_{k=1}^K D'_k \right\}, \left\{ \bigcup_{k=1}^K S_k \right\} \cap \left\{ \bigcap_{k=1}^K S'_k \right\} \right); \end{aligned} \quad (4.9)$$

$$\begin{aligned} (D_{wa}^{t*}, S_{wa}^{t*}) &= \left(\bigcap_{(t,c) \in \mathcal{E}(t)} D_{wa}^{tc}, \bigcup_{(t,c) \in \mathcal{E}(t)} S_{wa}^{tc} \right), \\ (D_{sa}^{t*}, S_{sa}^{t*}) &= \left(\bigcap_{(t,c) \in \mathcal{E}(t)} D_{sa}^{tc}, \bigcup_{(t,c) \in \mathcal{E}(t)} S_{sa}^{tc} \right). \end{aligned} \quad (4.10)$$

4.2 Simulations

We perform a simulation study to evaluate certain frequentist properties of each method for finding credible subgroup pairs. We are primarily concerned with the properties of our four different types of admissibility: weak and strong, each estimated via the fully adjusted and direct methods. Our operating characteristics of primary interest are the average sensitivity and specificity of the exclusive credible subgroup D under increasing numbers of endpoints and treatment arms.

Each simulated data set is produced with A arms, $N = 100$ patients per arm, K endpoints, and $P = 3$ covariates. For patient i in arm a , $\mathbf{x}_{ai} = (1, x_{ai2}, x_{ai3})$ is a prognostic covariate vector where x_{ai2} and x_{ai3} are discrete covariates randomly drawn from $\{-2, -1, 0, 1, 2\}$ with probabilities $\{1/16, 1/4, 3/8, 1/4, 1/16\}$, respectively. The same vector is used as the predictive covariate vector: $\mathbf{z}_{ai} = \mathbf{x}_{ai}$. The following model is used to produce the simulated data:

$$Y_{aik} | \eta_{aik}, \sigma_k^2 \sim \text{Normal}[\eta_{aik}, \sigma_k^2], \quad \eta_{aik} = \mathbf{x}'_{ai} \boldsymbol{\beta}_k + \mathbf{z}'_{ai} \boldsymbol{\gamma}_k^{(a)}, \quad (4.11)$$

where Y_{aik} is the response in the k th endpoint for patient i in arm a , $\beta_k \equiv (1, 1, 1)$ for all k , and the $\gamma_k^{(a)}$ are determined as follows: $\gamma_1^{(a)} = (0, 1/3, 0)$ for $1 < a < A$, $\gamma_1^{(A)} = (0, 1, 0)$, all other $\gamma_k^{(a)} = (0, 0, 0)$. The scenarios tested were $A = 2-8$ with $K = 1$, and $K = 1-8$ with $A = 2$. We simulated 1000 data sets per scenario, constructing 50% credible subgroup pairs.

The model used to fit the simulated data is the same, with vague priors $\sigma_k^2 \sim \text{InverseGamma}[10^{-4}, 10^{-4}]$, $\beta_{kp} \sim \text{Normal}[0, 10^4]$ for all k, p , and $\gamma_{kp}^{(1)} = 0$, $\gamma_{k1}^{(a)} \sim \text{Normal}[0, 10^4]$, and conservative interaction priors $\gamma_{kp}^{(a)} \sim \text{Normal}[0, 1]$ for $a > 1$ and all k, p . The model was fit using the NIMBLE R package [35] for 100 burn-in iterations and an additional 1000 recorded iterations for each simulated data set. Credible subgroups were constructed using (4.7).

We also compare the fully adjusted and direct methods to a “naive” method for determining admissibilities. We use the above regression model without treatment-covariate interactions to estimate an average treatment effect independently for each endpoint-treatment combination. For each draw from the joint posterior of the average treatment effects we compute draws of weak and strong admissibility, then use the posteriors of the admissibilities to make inferences at the 50% level. The direct method reduces to the naive method when there are no treatment-covariate interactions.

The results of the simulation study are displayed in Figure 4.1. In most cases, sensitivity falls and specificity remains high as the number of arms or endpoints increases, with the exception that the specificity of direct weak admissibility decreases as endpoints are added. Additionally, detection of strong admissibility is more difficult than detection of weak admissibility, and adjusting for multiplicity in the estimation of admissibilities (i.e. using the fully adjusted instead of direct method) reduces sensitivity. The naive approach retains very high sensitivity and very low specificity for weak admissibility in all presented scenarios, and very low sensitivity and very high specificity for strong admissibility in all presented scenarios.

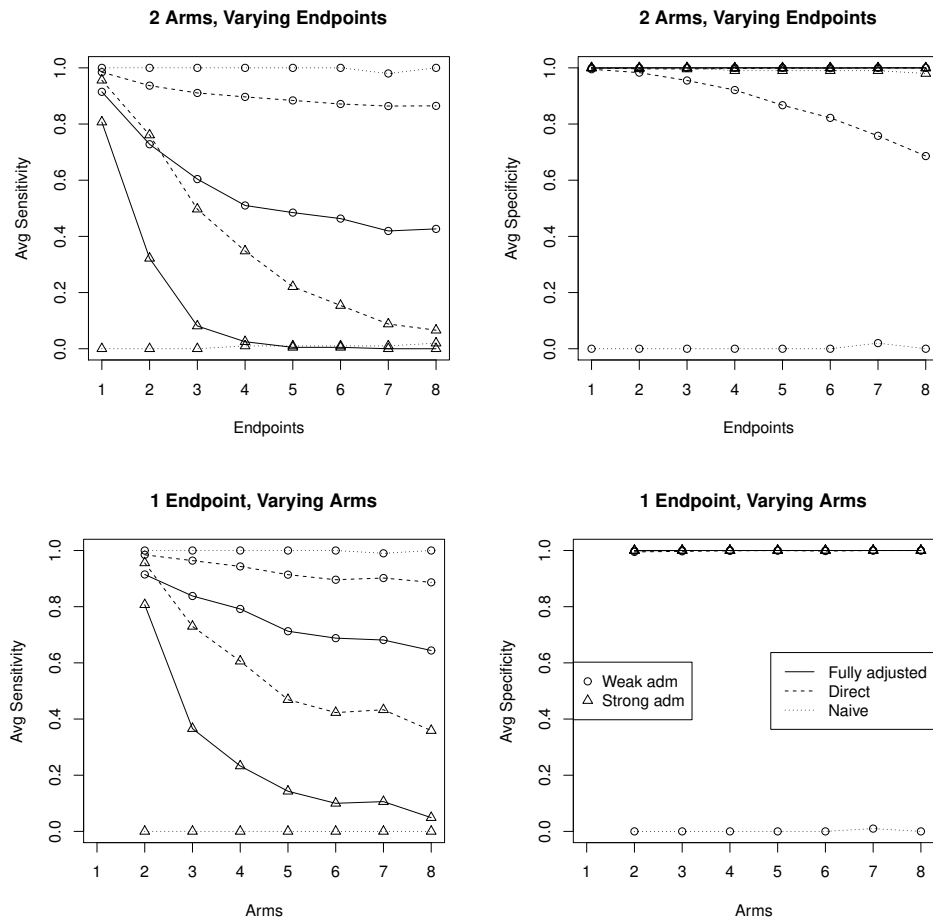


Figure 4.1: Simulated sensitivity (left column) and specificity (right column) for a study with $A = 2$ arms and varying number of endpoints (top row) and a study with $K = 1$ endpoints and a varying number of arms (bottom row). In most cases, sensitivity falls and specificity remains high as more arms or endpoints are added.

4.3 Analysis of multi-endpoint dataset

We illustrate the extended credible subgroups methods on the multi-endpoint Alzheimer’s disease dataset. Three doses (low, medium, high) of an experimental treatment are to be compared to active control and to a placebo. Baseline measurements for disease severity, age, sex, and carrier status of a genetic biomarker constitute covariates. After 24 weeks of treatment, two endpoints are of interest: improvement (negative change in disease severity) as the efficacy endpoint, and the reporting of at least one adverse event indicated by the attending physician to be possibly related to the treatment. The dataset includes a total of 331 patients across all arms. All covariates and the efficacy outcome are standardized for the analysis and displayed in their original units.

Let $a = 0, 1, 2, 3, 4$ denote the placebo, low, medium, and high doses of the test treatment, and active control treatment arms, respectively. For patient i , let Y_{ik} for $k = 1, 2$ denote the change in severity (continuous) and adverse event occurrence (binary) endpoints, respectively, and $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{z}_{i1} = \mathbf{z}_{i2}$ be the prognostic and predictive covariate vectors for each endpoint (including intercept, all considered as equal here). Let $\boldsymbol{\beta}_k$ be the vector of prognostic effects for the k th endpoint, and $\boldsymbol{\gamma}_k^{(a)}$ be the vector of predictive effects for the k th endpoint and treatment arm a , with $\boldsymbol{\gamma}_k^{(0)} = \mathbf{0}$. Also let $d^{(a)}$ be a scalar representing the level of activity of the drug dose in arm a compared to the maximum dose of the same drug, with $0 = d^{(0)} \leq d^{(1)} \leq d^{(2)} \leq d^{(3)} = d^{(4)} = 1$ and $\boldsymbol{\gamma}_k^{(1)} = \boldsymbol{\gamma}_k^{(2)} = \boldsymbol{\gamma}_k^{(3)}$, so that, for example, the effect of treatment $a = 2$ for a patient with predictive covariate vector \mathbf{z} is $d^{(2)}\mathbf{z}'\boldsymbol{\gamma}_k^{(2)}$. Assuming the outcomes are conditionally independent between patients, we use the endpoint likelihoods

$$Y_{i1}|\eta_{i1}, \sigma^2 \sim \text{Normal}[\eta_{i1}, \sigma^2], \quad Y_{i2}|\eta_{i2} \sim \text{Bernoulli}([\text{logit}^{-1}\eta_{i2}]), \quad (4.12)$$

with $\eta_{ik} = \mathbf{x}_{ik}^\top \boldsymbol{\beta}_{k*} + d^{(a_i)} \mathbf{z}_{ik}^\top \boldsymbol{\gamma}_{k*}^{(a_i)}$. We use the prior $\sigma^2 \sim \text{InverseGamma}[10^{-4}, 10^{-4}]$, $\boldsymbol{\beta}_{kp} \sim \text{Normal}[0, 10^4]$, $\boldsymbol{\gamma}_{k1}^{(a)} \sim \text{Normal}[0, 10^4]$ for $a > 0$, $\boldsymbol{\gamma}_{kp}^{(a)} \sim \text{Normal}[0, 1]$ for $a > 0$ and $p > 1$, $d^{(2)} \sim \text{Uniform}[0, 1]$, and $d^{(1)}|d^{(2)} \sim \text{Uniform}[0, d^{(2)}]$. Here we modestly shrink the treatment-covariate interactions toward zero to reflect the common prior belief that such interactions are usually small, and to obtain less variable estimates of conditional treatment effects; however we leave the priors for the prognostic effects and baseline treatment effect vague. As mentioned in the GLM example analysis, the Normal $[0, 1]$ prior is not as conservative in the logistic case as in the normal case,

though we leave the priors identical for illustration, as a more conservative prior on the treatment-covariate interactions for safety yield trivial credible subgroups ($S \setminus D = C$). A sensitivity analysis without any shrinkage did not yield qualitatively different results.

Before using our proposed methods, we analyze the data through a more standard approach. We use a Bayesian model and analysis, though with non-informative priors that correspond to a frequentist analysis. Because our aim is to discuss treatment-covariate interactions, which the study was not powered to detect, we decrease the nominal credible level to 50%. To make the approaches comparable, we will use the same credible level for our proposed analysis. All models are fit with 10,000 Gibbs sampler iterations after 1000 burn-in iterations. We first test the overall effects by removing all γ parameters from the model except the $\gamma_{k1}^{(a)}$, which then correspond to the overall treatment effects versus the placebo. In this analysis, there emerge significant overall efficacy differences between the active control and placebo, and between the test treatment and placebo. However, no significant safety differences nor an efficacy difference between the active control and the test treatment are uncovered.

We continue with a standard subgroup analysis, returning all γ parameters to the model and using minimally informative priors. At the 50% nominal credible level we find significant interactions between the test-placebo efficacy difference and all covariates; and between the test-placebo safety difference and baseline severity and age. We also find significant interactions between the test-active control efficacy difference and baseline severity and carrier status; and between the test-active control safety difference and sex and age. Using a Bonferroni-corrected α -level of $0.50/4$ to account for the four treatment-by-covariate interaction tests per treatment and endpoint (we aren't concerned with multiplicity of endpoints or treatments), we are left with only the interaction of the test-placebo efficacy difference with baseline severity and sex as significant.

We now estimate the average treatment effect of the test treatment versus the placebo in subgroups produced according to the significant interactions (post-Bonferroni) we identified. The treatment effect remains significant in a high-severity (> 22 , sample median) subgroup and a low-severity (≤ 22) subgroup, but when grouping by sex, there is a significant effect in males but not females. When the population is divided into four subgroups according to sex \times severity, both male subgroups and neither female subgroup show a significant effect. From this standard subgroup analysis, we get the

general idea that the male patients are the primary drivers of the treatment effect versus placebo. However, it is difficult to precisely determine who benefits from the treatment over the placebo, and especially what treatment effect exists between the test treatment and the active control.

We now compare the high dose test treatment to the placebo and active control simultaneously with respect to the weak and strong admissibility criteria, e.g. $\Delta_{wa}^{a*}(\mathbf{z}_i)$ and $\Delta_{sa}^{a*}(\mathbf{z}_i)$. In the former case, the benefiting subgroup is the population for which the test treatment is superior to both the placebo and active control with respect to at least one endpoint *or* is inferior to neither the placebo nor the active control with respect to either endpoint. In the latter case, the benefiting subgroup is the one for which the test treatment is superior to both the placebo and active control with respect to at least one endpoint *and* is inferior to neither the placebo nor the active control with respect to any endpoint. The criteria we select for superiority are a difference in expected change in disease severity of greater than $\delta_1 = 0$ and a log odds ratio of adverse event occurrence of less than $-\delta_2 = 0$. The criteria for noninferiority are a difference in expected change in disease severity of greater than $\epsilon_1 = -0.5$ (standard deviations of the response) and a log odds ratio of adverse event occurrence of less than $-\epsilon_2 = 0.18$ (corresponding to an odds ratio of approximately 1.20). Signs are switched for δ_2 and ϵ_2 because we want *reductions* in risk. The model is fit using 100,000 MCMC iterations after 10,000 burn-in iterations. Because of the high memory requirements of constructing credible subgroups over a continuous covariate space, every 10th iteration is used for the computation.

Figure 4.2 shows individual single-arm single-endpoint credible subgroups plots for each treatment-endpoint combination using $\alpha = 0.50$ for illustration. Because the thresholds for benefit differ from the thresholds for noninferiority, there are in fact two pairs of credible subgroups for each treatment-endpoint combination—one for benefit and one for noninferiority. Letting (D, S) and (D', S') be the pairs for benefit and noninferiority, respectively, we have $D \subseteq D' \subseteq S \subseteq S'$. The upper-left sub-figure shows that males with high disease severity tend to benefit from the test treatment versus the placebo, but in the bottom left sub-figure we detect more non-inferiority in female and low severity patients versus the active control. This hints that the active control and the test treatment may both favor male and high-severity patients relative to the placebo, but that the active control does so to a larger degree, perhaps due to more activity

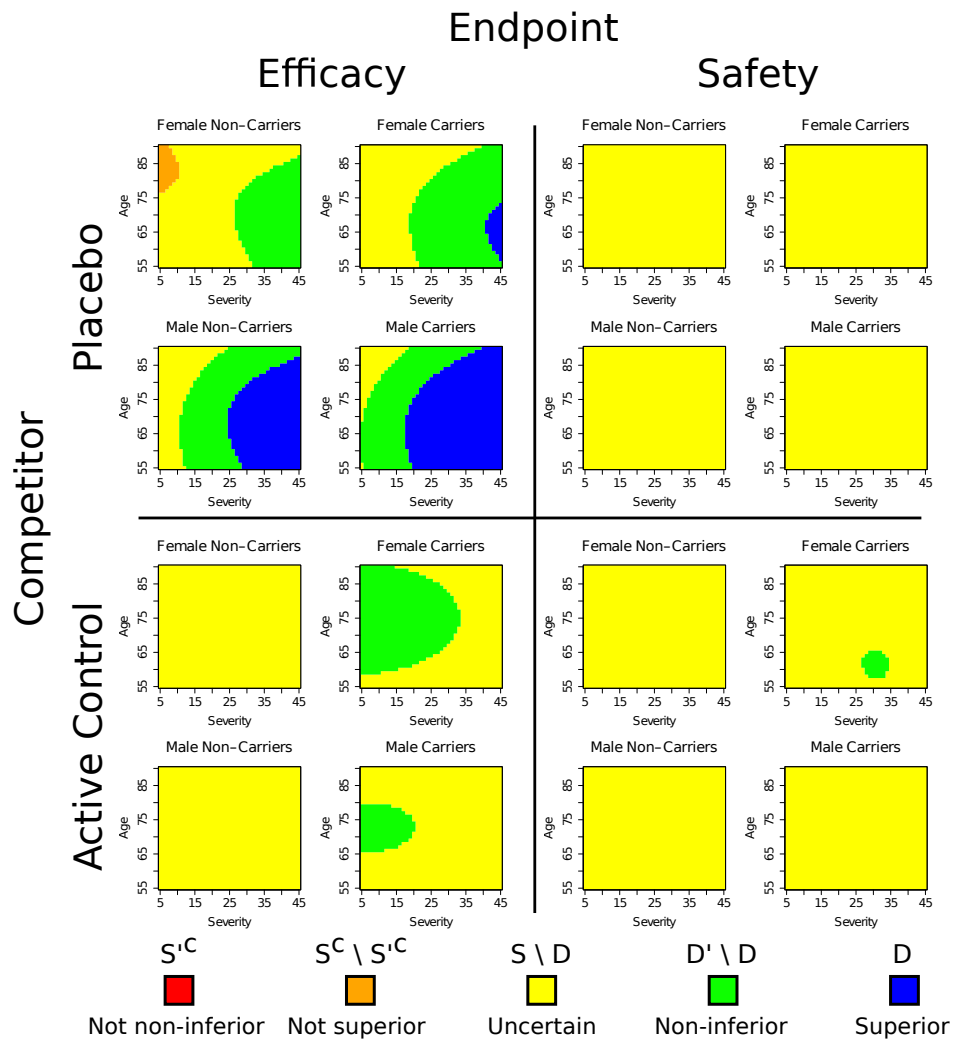


Figure 4.2: Credible subgroups for individual endpoint-competitor combinations at the 50% level.

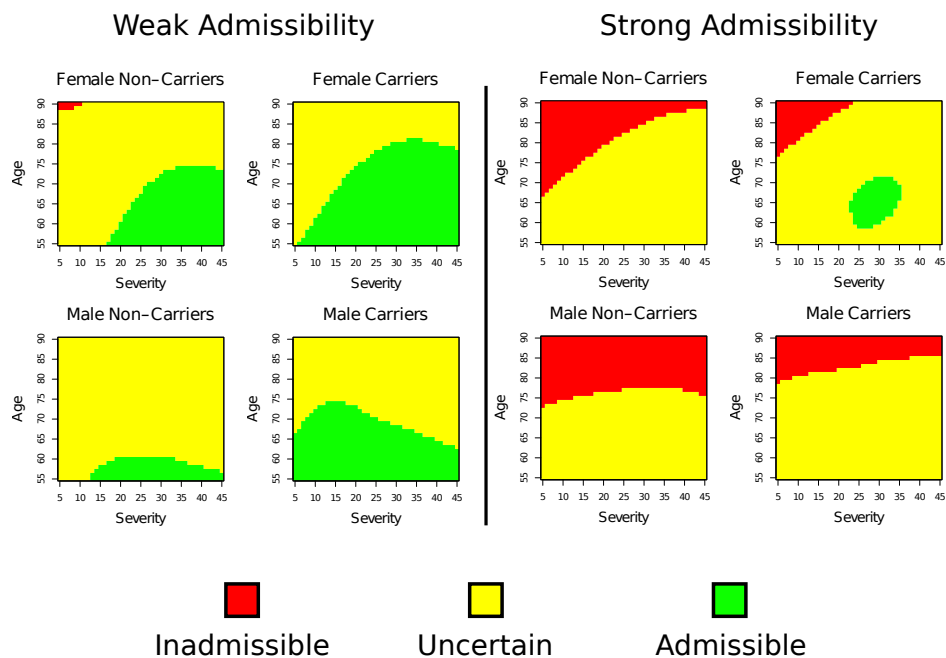


Figure 4.3: Credible subgroups for individual endpoint-competitor combinations at the 50% level.

of a similar biopharmaceutical mechanism. The right-hand side of the figure indicates mostly uncertainty in the relative safety profiles of the treatments, though it appears that female carriers are the most promising for non-inferiority to the active control.

Figure 4.3 shows credible subgroup pairs for weak and strong admissibility (via the direct methods) against both the placebo and active control. The left sub-figure shows that the exclusive credible subgroup for weak admissibility primarily contains younger patients, and is more present in females and carriers. The features of the weak admissibility credible subgroup plot appear (judging by Figure 4.2) to come primarily from the efficacy endpoint. The right sub-figure shows that the test treatment is not strongly admissible over an area generally opposite to that over which the test treatment is weakly admissible: older patients, especially males and non-carriers. Though the credible level used is too low to claim conclusive results (e.g., for a regulatory submission), the results provide evidence that the treatment effect is not homogeneous, and indicate which subgroups show promise for appropriately-powered studies in the future.

The model was also fit with spike-and-slab priors as in Section 3.2.2 for variable selection: the $\text{Normal}[0, 1]$ priors for the treatment-covariate interactions were exchanged for $\frac{1}{10}\text{Dirac}[0] + \frac{9}{10}\text{Normal}[0, 10^4]$ mixture distributions, where $\text{Dirac}[0]$ is a point mass at 0. The resulting individual credible subgroups by endpoint and arm as in Figure 4.2 exhibited much more homogeneity: the test treatment was superior to the placebo for all patients with respect to efficacy and indeterminate with respect to safety. Against the active control, the test treatment was noninferior with respect to efficacy for patients with baseline severity < 31 , and for women younger than 80 and men younger than 70 with respect to safety. The test treatment was weakly admissible for all patients, and strong admissibility was indeterminate for most patients, and negative for the patients with the highest baseline severity (near 45).

4.4 Discussion

The medical community recognizes the need to consider the characteristics of individual patients when deciding avenues of treatment. In addition to baseline covariates that are predictive of treatment effects with respect to single endpoints, it is also necessary to consider differences in individuals' preferences that may lead different patients to differentially value endpoints. For example, one patient may pursue the most efficacious treatment while another prefers a treatment with side effects that minimally affect quality of life.

The concept of admissibility provides a utility function-free approach to summarizing treatment effects with respect to multiple endpoints, and admits a natural extension to trials with more than two arms. In this paper we have also examined multiple definitions of admissibility in the clinical trial context, as well as estimators which do and do not adjust for the multiplicity of endpoints so that Type I error may be controlled. Finally, the credible subgroups method of earlier chapters provides a natural implementation for admissibility ideas by also adjusting for the multiplicity of covariate points, and we generalize an earlier method to handle settings outside of the normal linear model by requiring only a sample from the joint posterior of personalized treatment effects, allowing the consideration of generalized linear and other more sophisticated models.

While the confidence levels used in Figures 4.2 and 4.3 are too low for our results to be considered definitive, it is important to note that they are based on data from a study not powered to deliver simultaneous inference on multiple endpoints across arbitrary subgroups defined by up to four different covariates. So while these results are far from convincing for final regulatory approval, they do provide valuable information about the sort of enrollees that should be sought for future, more focused subgroup-confirmatory trials. For instance, the weak admissibility portion of Figure 4.3 suggests younger females with more severe dementia would make good candidates, whereas the strong admissibility portion discourages enrollment of older patients, particularly those with less severe dementia. Used in this way, our methods essentially become a useful tool for enrichment designs [32].

Finally, the relationship between identifying admissible treatments in the development and regulatory context treated here, and the single-patient focus of the dynamic treatment regime context, present an interesting duality between decisions made in relation to a given treatment versus a given patient. For example, developer-sponsored clinical trials may aim to secure regulatory approval for therapies in specific subpopulations, and optimal treatment regimes may subsequently be constructed on a per-patient basis from available treatments using the concepts of admissibility, which are similar to the non-domination criteria used in [47]. Attempts toward unifying development, regulatory, and patient-care contexts may represent a promising avenue for future research.

Chapter 5

Considerations for Practical Implementation

Previous chapters have focused primarily on developing the theory of credible subgroups, discussing implementation only through providing example analyses. This chapter more thoroughly addresses several, somewhat disjoint considerations that arise when implementing these methods in their originally intended scenario: design and analysis of clinical trials. Section 5.1 considers power computations for simple cases both *a priori* (for trial design) and *post hoc* for choosing the restriction of the covariate space of interest. Section 5.2 and Section 5.3 treat issues connected to final data analyses: Monte Carlo precision and diagnostics, respectively. Section 5.4 presents an option for reporting credible subgroups by building an easy-to-use calculator to determine the credible subgroup conclusion (or lack of one) for a given patient, aimed at clinical practitioners. Finally, Section 5.5 details essential software for the inference step of the credible subgroups approach, given a sample from the joint posterior of the personalized treatment effects or the parameters needed to compute them. Although the topics of most of these sections require substantial further investigation, it seems useful to record their states of development at this time.

5.1 Power computations

A crucial clinical trial design parameter that must be determined ahead of time is the sample size, which is typically selected to provide a targeted amount of power to detect a specified treatment effect while controlling Type I error. In planning trials for overall treatment effects, sample size estimations for a fixed power level generally boil down to deciding on a clinically relevant effect size Δ and a sensible corresponding outcome variance σ^2 . We will not consider complications such as time-to-event endpoints, drop-out, etc.

Benefiting subgroup identification trials immediately present the complication that Δ is not assumed to be constant across the covariate space. While the minimum clinically relevant effect size may be constant across the covariate space, it must be kept in mind that the estimated personalized treatment effect at one covariate point may be strongly affected by the true value of the PTE at nearby points, and the shape of the PTE surface affects the dependency of the posterior distribution of the PTE across the covariate space, and thus $W_{\alpha,C}^*$. However, absent specific prior information about how the PTE might vary, a constant hypothesized PTE seems a reasonable starting point.

The distribution of covariates in the study sample also has a greater effect on credible subgroups analyses than on tests for overall treatment effects. This dependency turns out to be part headache and part opportunity: if the covariate distribution is not controlled at enrollment, an understanding of the population covariate distribution and any enrollment biases is essential for accurate power calculations; however, controlling the covariate distribution through targeted enrollment, quotas, or enrichment designs can target specific populations while the credible subgroups methods provide a principled approach to determining the extent of generalizability.

5.1.1 Local power

Given a covariate space of interest, a sample of covariate points, and a hypothesized PTE function Δ_0 , and under the assumption of asymptotic normality of the joint posterior of the $\Delta(\mathbf{z})$, Algorithm 4 may be used to estimate the *local power* (power to detect an effect of size $\Delta_0(\mathbf{z})$ at \mathbf{z}).

The local power estimate is derived by assuming asymptotic joint normality of

Algorithm 4 Local (conditioning on \mathbf{x}) Power Simulation

- 1 Simulate outcomes for patients under the assumed Δ_0 ;
 - 2 Fit planned model and compute simultaneous credible band (3.3), including estimating $W_{\alpha,C}^*$;
 - 3 Estimate posterior standard errors $\text{SE}\{\Delta(\mathbf{z})\}$ for PTEs;
 - 4 Estimate power as $1 - \Phi \left[W_{\alpha,C}^* - \Delta_0(\mathbf{z}) / \text{SE}\{\Delta(\mathbf{z})\} \right]$, where Φ is the standard normal distribution function and $\Delta_0(\mathbf{z})$ is the hypothesized effect to be detected.
-

$\bar{\Delta}(\mathbf{z})$, the posterior mean of the PTE at \mathbf{z} , and evaluating the frequentist probability $\text{P} \left[\bar{\Delta}(\mathbf{z}) - W_{\alpha,C}^* \sqrt{\text{Var} [\bar{\Delta}(\mathbf{z})]} > \delta \mid \Delta(\mathbf{z}) = \Delta_0(\mathbf{z}) \right]$ with the intention of determining the power to detect benefit. Other PTE surfaces may be assumed and power estimated for them, but without prior information, the constant surface computations are perhaps most easily interpreted. Additionally, the above algorithm may depend on nuisance parameters such as error variances. It may be possible to estimate these nuisance parameters through empirical Bayesian methods (e.g., using restricted maximum likelihood to estimate the hyperparameters from their marginal distributions) using the observed data without introducing significant bias.

Although simulations of multiple sample covariate distributions and hypothesized PTE surfaces are still necessary, Algorithm 4 negates the need to simulate multiple sets of outcomes for each configuration, and can be used after data collection to choose the restriction of the covariate space to be tested, as in the following subsection. Note also that the algorithm yields a conservative power estimate because it is based on the single-step procedure rather than the step-down one.

5.1.2 Choice of covariate space

The example analysis in Section 2.3, specifically the contrived 50% credible subgroups, highlighted a problem in which inferences could be made that were valid within the model but were dubious in practice due to being outside the portion of the covariate space with observations and common support between arms. The semiparametric analysis of Section 3.3.2 illustrated the rapid increase in the standard error of the PTE

estimates from the penalized spline model, which can be interpreted as partially addressing the observation density and common support problems [49]. However, relying on high standard errors to address the support problems can result in excessively conservative inferences due to the inclusion of unnecessary covariate points in the restricted covariate space.

A combination of conditional local power simulations and direct examination of the observation density and common support in the covariate space may be used to choose the appropriate restriction of the covariate space for analysis. Figure 5.1 displays the sample distribution of covariate points and shades the region over which the 95% credible subgroups have at least 5% power to detect a uniform benefit of 1 standard deviation (5 points). Restricting severity to $[9, 49]$ and d-rate to $[0, 8]$ we include this entire region and exclude less than 9% of patients, while reducing the size of the covariate space by more than half. It can be seen that observations, and especially observations from different treatment arms, are hopelessly sparse in much of the excluded region. The restricted region was used in the example analysis in Section 3.3.2 and yielded an exclusive credible subgroup with approximately 17% more cells than the corresponding exclusive credible subgroup when the full observed ranges of severity and d-rate are used.

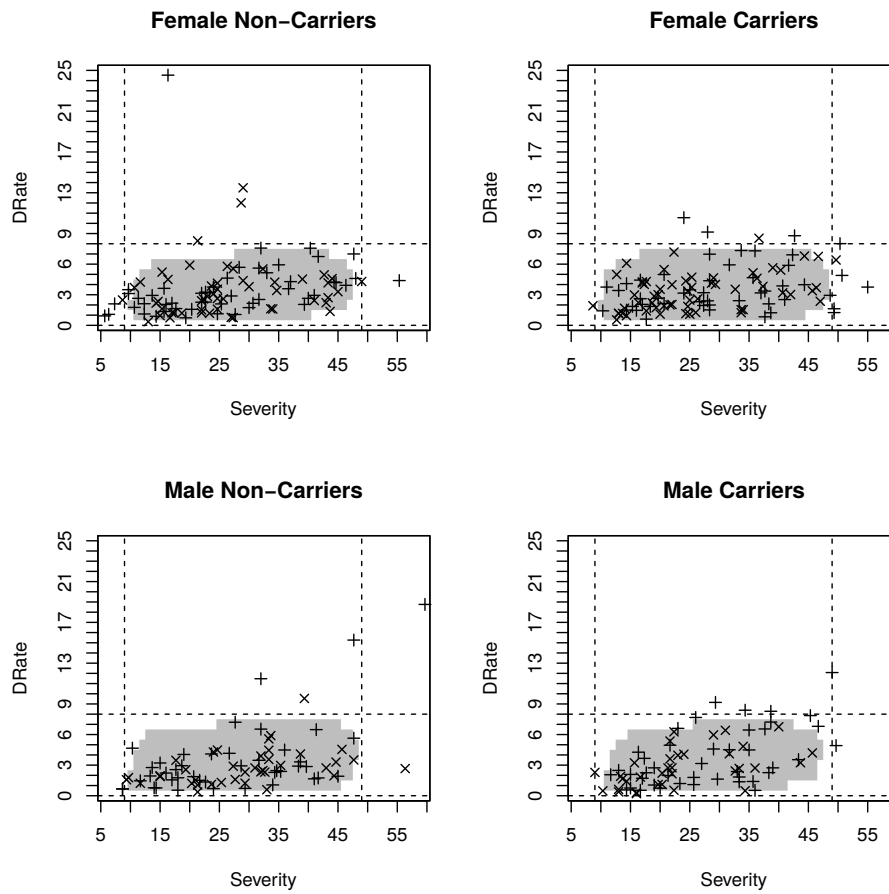


Figure 5.1: The shaded region is the region for which the power to detect a 1 standard deviation (5-point) benefit at the 95% credible level is at least 5% when the entire empirical covariate space is used. Patients receiving placebo are represented by \times , and those receiving the standard of care by $+$.

5.2 Monte Carlo precision

Several aspects of credible subgroup construction conspire to make estimation of Monte Carlo precision challenging. First, the variability of credible subgroup estimates (D, S) cannot be directly described in terms of standard errors: unlike scalar estimators, they vary in terms of covariate points being included in or excluded from constructed subgroups. Second, the bounds on the personalized treatment effect at each covariate point used to construct the credible subgroups are functions of three Monte Carlo-estimated quantities with complex dependencies, one a maximum of a large but fixed number of dependent variables. Third, there are no apparent dramatic shortcuts to implementing computational techniques for estimating Monte Carlo variability, such as cheap jackknife [50, 51, 52] or bootstrap [53] estimators. Finally, the step-down procedure adds complexity on top of that already present in estimating the Monte Carlo variability of the single-step procedure. In this section we discuss approximations that can be used in estimating the Monte Carlo variability of the single-step procedure. The discussion will explicitly treat the exclusive credible subgroup, though all points apply to the inclusive credible subgroup as well, with straightforward modifications. Throughout, all distributions are in the posterior, with conditioning on observed data suppressed in the notation.

The single-step classification of a covariate point \mathbf{z} with respect to the exclusive credible subgroup is determined entirely by the posterior quantity

$$L(\mathbf{z}) = \bar{\Delta}(\mathbf{z}) - W_{\alpha, C}^* \sqrt{\text{Var}[\Delta(\mathbf{z})]}, \quad (5.1)$$

where $\bar{\Delta}(\mathbf{z})$ and $\text{Var}[\Delta(\mathbf{z})]$ are the posterior mean and variance of $\Delta(\mathbf{z})$, and $W_{\alpha, C}^*$ is the $1 - \alpha$ quantile of

$$W_C = \sup_{\mathbf{z} \in C} \frac{|\Delta(\mathbf{z}) - \bar{\Delta}(\mathbf{z})|}{\sqrt{\text{Var}[\Delta(\mathbf{z})]}}. \quad (5.2)$$

Given a finite, discrete covariate space C and a sample of size M from the joint posterior

of $\mathbf{\Delta}(\mathbf{C})$ indexed by m , define the Monte Carlo estimates

$$\begin{aligned}
\widehat{\Delta}(\mathbf{z}) &= \frac{1}{M} \sum_{m=1}^M \Delta^{(m)}(\mathbf{z}), \\
\widehat{\text{Var}}[\Delta(\mathbf{z})] &= \frac{1}{M-1} \sum_{m=1}^M \left[\Delta^{(m)}(\mathbf{z}) - \widehat{\Delta}(\mathbf{z}) \right]^2, \\
\widehat{W}_{\mathbf{C}}^{(m)} &= \max_{\mathbf{z} \in \mathbf{C}} \frac{\left| \Delta^{(m)}(\mathbf{z}) - \widehat{\Delta}(\mathbf{z}) \right|}{\sqrt{\widehat{\text{Var}}[\Delta(\mathbf{z})]}}, \\
\widehat{W}_{\alpha, \mathbf{C}}^* &= \min_{1 \leq m \leq M} \left\{ \widehat{W}_{\mathbf{C}}^{(m)} : 1 - \alpha \leq \frac{1}{M} \sum_{l=1}^M \mathbb{I} \left(\widehat{W}_{\mathbf{C}}^{(l)} \leq \widehat{W}_{\mathbf{C}}^{(m)} \right) \right\}.
\end{aligned} \tag{5.3}$$

Then (5.1) can be approximated by the Monte Carlo estimate

$$\widehat{L}(\mathbf{z}) = \widehat{\Delta}(\mathbf{z}) - \widehat{W}_{\alpha, \mathbf{C}}^* \sqrt{\widehat{\text{Var}}[\Delta(\mathbf{z})]}.$$
 \tag{5.4}

Note that when $\mathbf{\Delta}(\mathbf{C})$ is asymptotically multivariate normal, the Monte Carlo variances of $\widehat{\Delta}(\mathbf{z})$ and $\widehat{\text{Var}}[\Delta(\mathbf{z})]$ have analytical expressions. However, that of $\widehat{W}_{\mathbf{C}}^{(m)}$ and therefore that of $\widehat{W}_{\alpha, \mathbf{C}}^*$ remain intractable.

The intractability of $\widehat{W}_{\alpha, \mathbf{C}}^*$ suggests using further Monte Carlo methods to estimate its variance. However, the resources spent performing such a simulation may be better spent simply drawing a larger posterior sample to increase precision, unless doing so is substantially more expensive. One possible approach to lowering the cost of estimating the Monte Carlo variance of $\widehat{L}(\mathbf{z})$ is to assume loose dependence of $\widehat{W}_{\alpha, \mathbf{C}}^*$ on $\widehat{\Delta}(\mathbf{z})$ and $\widehat{\text{Var}}[\Delta(\mathbf{z})]$ for each \mathbf{z} individually, and then use $\widetilde{V} = \text{Var} \left[\widehat{W}_{\alpha, \mathbf{C}}^* \mid \widehat{\Delta}(\mathbf{z}), \widehat{\text{Var}}[\Delta(\mathbf{z})] \right]$ as an approximation to $\text{Var} \left[\widehat{W}_{\alpha, \mathbf{C}}^* \right]$. The conditional variance may be estimated cheaply by resampling the $\widehat{W}_{\mathbf{C}}^{(m)}$ directly, instead of the $\Delta^{(m)}(\mathbf{z})$. We then have

$$\begin{aligned}
\text{Var} \left[\widehat{L}(\mathbf{z}) \right] &= \text{Var} \left[\widehat{\Delta}(\mathbf{z}) - \widehat{W}_{\alpha, \mathbf{C}}^* \sqrt{\widehat{\text{Var}}[\Delta(\mathbf{z})]} \right], \\
&\approx \text{Var} \left[\widehat{\Delta}(\mathbf{z}) \right] + \text{Var} \left[\widehat{W}_{\alpha, \mathbf{C}}^* \right] \text{Var} \left[\sqrt{\widehat{\text{Var}}[\Delta(\mathbf{z})]} \right] \\
&\quad + \text{Var} \left[\widehat{W}_{\alpha, \mathbf{C}}^* \right] \text{E} \left[\sqrt{\widehat{\text{Var}}[\Delta(\mathbf{z})]} \right]^2 + \text{E} \left[\widehat{W}_{\alpha, \mathbf{C}}^* \right]^2 \text{Var} \left[\sqrt{\widehat{\text{Var}}[\Delta(\mathbf{z})]} \right], \\
&\approx \frac{1}{M} \widehat{\text{Var}}[\Delta(\mathbf{z})] + \widetilde{V} \frac{\widehat{\text{Var}}[\Delta(\mathbf{z})]}{2(M-1)} + \widetilde{V} \widehat{\text{Var}}[\Delta(\mathbf{z})] + \left(\widehat{W}_{\alpha, \mathbf{C}}^* \right)^2 \frac{\widehat{\text{Var}}[\Delta(\mathbf{z})]}{2(M-1)}.
\end{aligned} \tag{5.5}$$

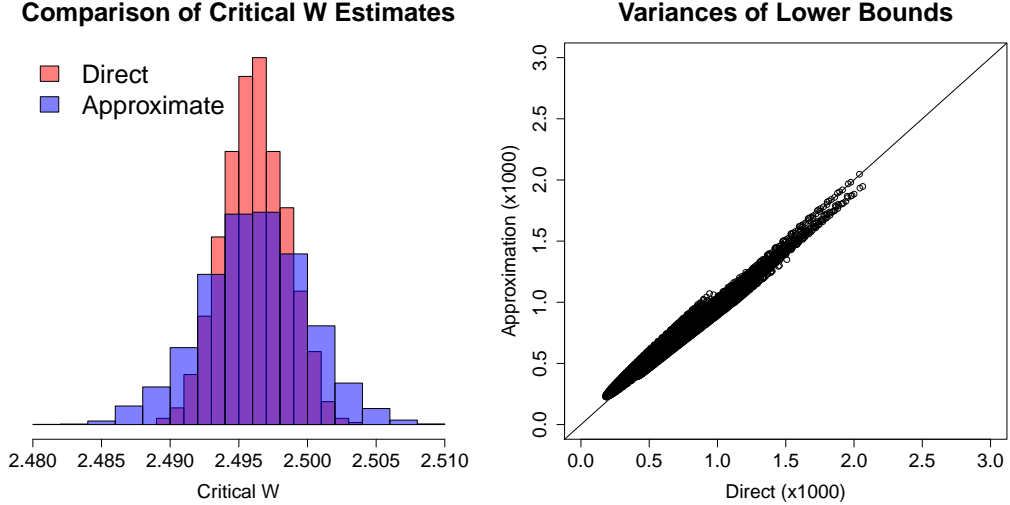


Figure 5.2: Direct and mean-aligned conditional bootstrap sample of $\widehat{W}_{\alpha,C}^*$ (left) and direct and conditional approximations of $\widehat{L}(z)$ (right).

To evaluate this approximation, we produced 1000 samples of 100,000 independent draws from the posterior distribution of the normal linear model fit in Section 2.3, and for each sample computed all $\widehat{\Delta}(z)$, $\widehat{\text{Var}}[\Delta(z)]$, and $\widehat{W}_{\alpha,C}^*$. We also drew 10,000 bootstrap samples of $\widehat{W}_{\alpha,C}^*$ from the $\widehat{W}_C^{(m)}$ produced from one of the above samples, holding $\widehat{\Delta}(z)$ and $\widehat{\text{Var}}[\Delta(z)]$ fixed. The correlation between the $\widehat{\Delta}(z)$ and $\widehat{W}_{\alpha,C}^*$ was very small (range -0.07–0.07), and between $\widehat{\text{Var}}[\Delta(z)]$ and $\widehat{W}_{\alpha,C}^*$ was approximately 0.33 (range 0.26–0.41), and we expect the dependency to weaken with the dependence among the $\Delta(z)$ in, e.g., semiparametric models. Figure 5.2 compares the bootstrapped conditional approximation to the true Monte Carlo distribution of $\widehat{W}_{\alpha,C}^*$ (left), and the approximated variances of all the $\widehat{L}(z)$ to their true counterparts (right). In this example the conditional approximation for the variance of $\widehat{W}_{\alpha,C}^*$ is conservative and those of the $\widehat{L}(z)$ appear adequate. Using the asymptotic conditional variance of $\widehat{W}_{\alpha,C}^*$, $\alpha(1-\alpha)/[M\widehat{f}(\widehat{W}_{\alpha,C}^*)]$, where \widehat{f} is the Gaussian kernel density estimate using the Sheather & Jones [54] bandwidth, in this case yields a less conservative estimate of $\text{Var}[\widehat{W}_{\alpha,C}^*]$ but a slightly less accurate estimate of $\text{Var}[\widehat{L}(z)]$.

Once Monte Carlo variances of the lower bounds have been estimated, they may be used to check which covariate points may change membership status in the exclusive

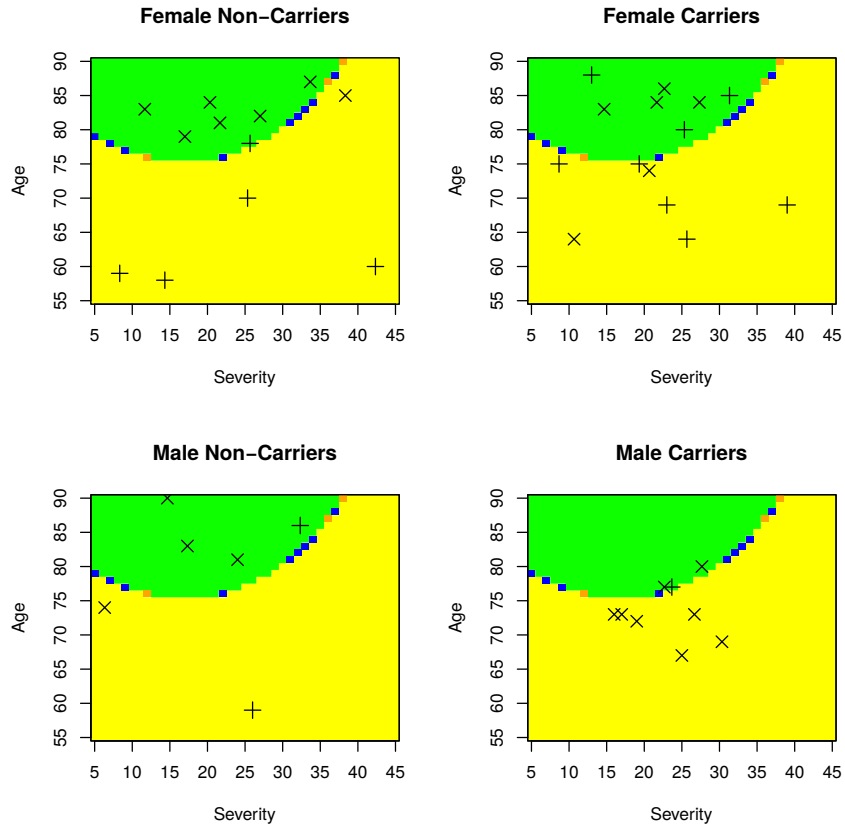


Figure 5.3: Uncertainty in credible subgroups. Orange points are part of the estimated exclusive credible subgroup but not part of the 0.5th percentile of exclusive credible subgroups, and blue points are not part of the estimated exclusive credible subgroup but are part of the 99.5th percentile of exclusive credible subgroups.

credible subgroup. Figure 5.3 marks the covariate points whose membership changes when using the endpoints of the 99% Monte Carlo confidence intervals for the lower bounds instead of the original estimates. As may be expected, these points lie at the boundary of the exclusive credible subgroup. It is likely very difficult to draw a large enough posterior sample to avoid such variation, even at modest levels of confidence, when the boundary of a credible subgroup contains many points along high-resolution continuous covariates, as by definition the upper and lower simultaneous credible bands at the boundary are very close to the threshold between benefit and lack of it.

5.3 Diagnostics

We have found useful several diagnostic tools for the credible subgroups approach. Many have already been exhibited in the example analyses of previous sections, though we include them in the following list for completeness.

1. Standard diagnostic tools for the underlying regression model, to evaluate model fit;
2. Histograms and quantile plots of personalized treatment effect posteriors to evaluate normality assumptions, if applicable;
3. Trace plots of the same, to evaluate Markov chain Monte Carlo convergence, if applicable;
4. Decomposition of credible subgroups into posterior means and standard deviations of personalized treatment effects, to determine whether failure to classify covariate points into D or S^c is the result of neutral treatment effects estimates or high standard errors;
5. Posterior mean–standard deviation plots (or *funnel plots*), to show trends between estimates and standard errors as they relate to credible subgroups;
6. Comparison of sample covariate distribution to credible subgroups and decompositions, to detect uncomfortable extrapolations;
7. Maximum credible level computations, to give a fuller picture of possible credible subgroups;
8. Plots showing Monte Carlo uncertainty in the credible subgroups construction, to determine whether they are tolerable.

The tool which is neither previously displayed nor obvious in construction and interpretation is the funnel plot. The remainder of this section describes it in some detail.

Since the credible subgroups are determined by comparing the posterior quantities $\bar{\Delta}(\mathbf{z}) \pm W_{\alpha,C}^* \sqrt{\text{Var}[\Delta(\mathbf{z})]}$ against the treatment effect threshold δ , with $W_{\alpha,C}^*$ possibly determined by the step-down procedure, we can plot $\sqrt{\text{Var}[\Delta(\mathbf{z})]}$ versus $\bar{\Delta}(\mathbf{z}) - \delta$

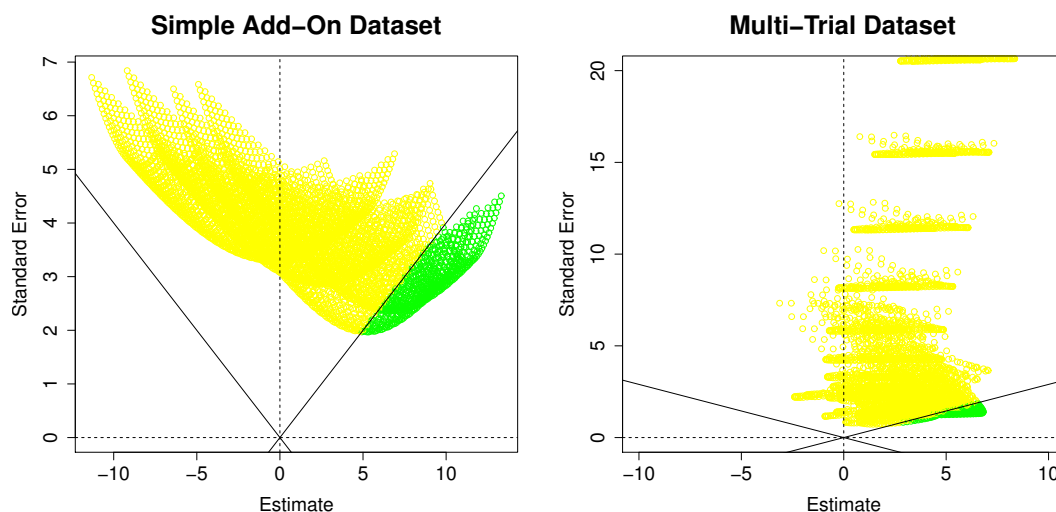


Figure 5.4: Funnel plots for two previously analyzed datasets. Green points are members of the exclusive credible subgroup, and yellow points are part members of the inclusive but not exclusive credible subgroup.

across $z \in C$, and determine points' credible subgroup memberships by examining their positions in relation to the lines through the origin with slopes $\pm 1/W_{\alpha,C}^*$. While the spatial information among the covariate points is lost, such a plot can be produced regardless of the dimension of the covariate space and the number of levels of each covariate.

Figure 5.4 presents the funnel plots for two previously analyzed datasets: the simple add-on dataset analyzed with the basic linear model in Section 2.3, and the multi-trial dataset analyzed with the additive penalized cubic spline model in Section 3.3.2. As also shown by the contour plots in Figure 2.4, the lower personalized treatment effect estimates in the first dataset correspond to higher posterior standard deviations. At least two points are deducible from both the contour and mean-standard deviation plots, but are more apparent from the latter:

1. For these data, the reliability of estimates below the treatment effect threshold is generally lower than that of those above the threshold, thus as more data are collected so that the standard errors decrease everywhere, points on the plot tend to be more likely to cross the vertical axis from left to right than the reverse;

2. As the credible level is lowered (i.e., the sloped lines approach the vertical axis), many more points join the exclusive credible subgroup before any exit the inclusive credible subgroup.

The plot corresponding to the multi-trial dataset consists of a dense region in which the standard errors are mostly uncorrelated with the estimates, and bands of relatively constant high standard errors over a wide range of estimates. These bands correspond to edges of the covariate space and continue well above standard errors of 20 units. In this case we can be even more confident that most covariate points have a positive PTE, which in order to detect we must collect more data.

5.4 Reporting

Throughout the previous chapters, example analyses investigated the personalized treatment effect in relation to two continuous and two binary predictive covariates. As a result, we were able to present the credible subgroups graphically. Such graphical summaries are likely possible if the only predictive covariates used are the standard demographic variables sex, age, race, and ethnicity. When more than two continuous covariates are predictive, or when there are a large number of categorical predictive covariates, such graphical displays are likely impossible or too cumbersome to produce, disseminate, and use.

One way of presenting results of credible subgroups analyses with (almost) arbitrarily many covariates and levels is via a calculator. When the threshold defining clinically significant benefit is fixed, the maximum credible level computation of Algorithm 3, along with a data frame representing the corresponding covariate space, may be used to produce a calculator such as that shown in Figure 5.5. The calculator presented was produced using the R package `shiny` [55], and can be deployed as a webpage. Software for producing such a calculator is included in Appendix C, and its usage demonstrated in Section 5.5. Other options include locally hosted computer or smartphone applications. Such a calculator would enable a clinician or other interested party to input a real or hypothetical patient's predictive covariate profile and receive an explanation of what determination can be made about that covariate profile, and at what credible level it can be made. To prevent misuse, a lower limit on the credible level may be implemented

Credible Subgroups Calculator

Baseline Severity

27 ▼

Rate of Decline

5 ▼

Sex

M ▼

Carrier Status

NON-CARRIER ▼

Result

A conclusion of **benefit** for patients with the above predictive covariate profile may be made at a maximum credible level of **92.38%**. At higher credible levels, no conclusion may be drawn.

Figure 5.5: Example of calculator for reporting credible subgroup results, using the results of the multi-trial dataset analysis.

so that if the maximum credible level falls below it, the output does not include whether a lower credible level would result in a conclusion of benefit or no benefit.

Such calculators are in fact simply user-friendly front-ends to large tables, which could in theory be presented statically. For example, the table of maximum credible levels for the analysis of the simple add-on therapy dataset presented in Section 2.3 would need to contain entries corresponding to 5904 covariate points, determined by four factors. Whether presented in tabular or electronic form, the maximum credible levels presented could be interpreted as a predictive “score” for the purposes of drug labeling, which is used to advise practitioners on prescribing medications. The label summary could include a statement that recommends use in populations with sufficiently high predictive scores (the maximum credible levels) by defining a threshold and directing prescribers to a full table or online calculator, provided by the FDA, for computing the score for an individual patient. While this arrangement would be more complex than labels currently used, precedent exists in using risk scores to guide treatment decisions by practitioners. For example, the Canadian Cardiovascular Society advocates using the Framingham Risk Score [56] for cardiovascular disease to stratify patients into groups for which different therapies are recommended [57]. While the Framingham Risk Score is linear in its component risk factors and can therefore be computed by adding up “points” associated with different levels of those individual risk factors, full risk tables and online calculators are often used, especially for adjusted variants of the score.

If more concise guidelines are a high priority, modeling choices can be made that yield more easily expressible credible subgroups. For example, using a variable selection method and conditioning on one or a few of the most probably models, as in Section 3.2.2, may yield credible subgroups that depend only on categorical covariate, or at most one continuous covariate. However, in the interest of principled inference, we recommend avoiding modeling choices that serve only to make summaries shorter, and if such simplified summaries are needed to build and communicate intuition, they should be presented as such with the caveat that a calculator should be used for final decisions.

5.5 Software

This section describes the use of several R functions that, given a description of the covariate space and a sample from the posterior distribution of the posterior treatment effects, produce credible subgroups and a calculator as described in Section 5.4. The full code is reproduced in Appendix C.

The three functions related to the construction of credible subgroups themselves are

- `sim.cred.band`, which constructs a restricted covariate space simultaneous credible band for the PTEs;
- `credsubs`, which determines the membership of each covariate point in the exclusive and inclusive credible subgroups using the sequential testing procedure described by Algorithm 2;
- `credsubs.level`, which determines via Algorithm 3 the maximum credible level at which each covariate point is not a member of $S \setminus D$, and attaches an attribute "sign" which is 1 for covariate points in D and -1 for those in S^c at that level.

Each of the above functions relies on the same three major input objects:

- `params`, a matrix whose rows are draws from the joint posterior of the PTEs, or of the parameters necessary to compute them;
- `design`, a matrix whose rows represent the points in the covariate space of interest;
- `FUN`, a function which takes as arguments a row of `design` and the entirety of `params` and returns the corresponding sample from the posterior of the PTE at the covariate point corresponding to that row of `design`.

By default, `design` is set to `NULL`, in which case `params` is taken to be a sample from the joint posterior of the PTEs and `FUN` is ignored. Also by default, `FUN` is set to `function(x, params) { params %*% t(x) }` so that a linear model is assumed whenever `design` is provided. The outputs are aligned with the rows of `design` if not `NULL`, and with the columns of `params` otherwise.

	<code>sim.cred.band</code>	<code>credsubs</code>	<code>credsubs.level</code>
<code>cred.level</code>	✓	✓	
<code>threshold</code>		✓	✓
<code>method</code>	✓	✓	✓
<code>step.down</code>		✓	✓

Table 5.1: Applicability of behavior arguments to primary functions.

Four arguments determine the behavior of the above functions in ways that may affect the primary outputs:

- `cred.level`, the credible level $1 - \alpha$ (default 0.95) at which simultaneous credible bands or credible subgroups are constructed;
- `threshold`, the threshold δ (default 0) above which a PTE is considered clinically beneficial;
- `method`, either "asymptotic" (default) or "quantile", specifying which simultaneous credible band construction is used;
- `step.down`, whether or not the step-down procedure should be used (default `TRUE`).

Table 5.1 displays the applicability of each of the above arguments to the primary functions.

A fourth function, `build.shiny.object`, packages and saves an R data file containing the output of `credsubs.level` and a formatted description of the covariate space `cov.space`, which may then be automatically loaded by the `shiny` app provided to produce a calculator as in Section 5.4 by placing it in the same directory as the app. The distinction between `design` and `cov.space` is that `design` is a matrix intended to be used in numerical computations while `cov.space` is a data frame meant to describe the covariate space in a human-readable form. For example, it is easiest to represent factors in `design` according to whichever parameterization was used in the regression model that produced `params` (e.g., dummy variables), but factors in `cov.space` should be `factor` objects so that textual level names are displayed.

5.5.1 A nonparametric example

We first present an example using a nonparametric (BART) fit of the multi-trial dataset. This displays the simplest usage of the software.

Assuming the credible subgroups functions have already been loaded, we proceed by loading the `BayesTree` package, which fits the BART model, and the data, then setting the random seed:

```
require(BayesTree)
data <- read.csv("data-simplified.csv")
set.seed(1)
```

Next, we build the training data and test set for BART. The `bart` function uses `x.train` as the covariates and `y.train` as the observed response to fit the model, then includes in its output `yhat.test`, a matrix whose rows are posterior estimates at covariate points described by `x.test`, which can be thought of as our `design` matrix. Because `bart` naturally handles factors, we do not need to reparameterize them.

```
x.train <- data[, c("Treatment",
                  "Baseline.Severity", "Rate.of.Decline",
                  "Sex", "Carrier.Status")]
y.train <- data[, "Improvement"]

design <- expand.grid("Treatment"=factor(c("Placebo", "Standard of Care")),
                  "Baseline.Severity"=9:49,
                  "Rate.of.Decline"=0:8,
                  "Sex"=factor(c("F", "M")),
                  "Carrier.Status"=factor(c("NON-CARRIER", "CARRIER")))
```

Here the design matrix has contains two rows for each covariate point: one in each treatment arm. The covariate points form a regular grid across the restricted covariate space, with baseline severity and rate of decline changing in increments of one.

```
> head(design)
      Treatment Baseline.Severity Rate.of.Decline Sex Carrier.Status
1      Placebo           9             0 F      NON-CARRIER
2 Standard of Care           9             0 F      NON-CARRIER
3      Placebo          10             0 F      NON-CARRIER
4 Standard of Care          10             0 F      NON-CARRIER
5      Placebo          11             0 F      NON-CARRIER
6 Standard of Care          11             0 F      NON-CARRIER
```

We can then fit the BART model and produce our `params` object. Since `bart` outputs two predictions at every covariate point, we compute the PTE by subtracting the prediction at each covariate point with `Treatment` equal to "Placebo" from the prediction of the corresponding covariate point with `Treatment` equal to "Standard of Care".

```
fit <- bart(x.train=x.train, y.train=y.train, x.test=design,
           ndpost=10000, nskip=1000, printevery=1000)

predictctl <- bart$yhat.test[, design$Treatment == "Placebo"]
predicttrt <- bart$yhat.test[, design$Treatment == "Standard of Care"]

params <- predict.trt - predictctl
rm(predict.trt, predictctl)
```

Since the `params` object is a matrix of posterior PTE draws, we do not need to pass `design` or `FUN` to any of the credible subgroups functions:

```
scb <- sim.cred.band(params=params, cred.level=0.80)
cs <- credsubs(params=params, cred.level=0.80)
csl <- credsubs.level(params=params)
```

The primary outputs of each function are two vectors aligned with the columns of `params`, representing a result at each of the covariate points:


```

> head(scb$lower)
[1] -2.916278 -2.803885 -2.746283 -2.643345 -2.585534 -2.523565
> head(scb$upper)
[1] 5.299417 5.215144 5.188561 5.108275 5.034181 4.982117
> head(cs$exclusive)
[1] FALSE FALSE FALSE FALSE FALSE FALSE
> head(cs$inclusive)
[1] TRUE TRUE TRUE TRUE TRUE TRUE
> head(csl)
[1] 0.2324 0.2324 0.2324 0.2324 0.2324 0.2324
> head(attr(csl, "sign"))
[1] 1 1 1 1 1 1

```

Finally, to produce the data file used by the calculator, we build the `cov.space` object, which here differs from the `design` object in that the `Treatment` variable is not included and the column names do not have to exactly match those in the dataset.

```

cov.space <- expand.grid("Baseline Severity"=9:49,
                        "Rate of Decline"=0:8,
                        "Sex"=factor(c("F", "M")),
                        "Carrier Status"=factor(c("NON-CARRIER", "CARRIER")))

build.shiny.object(csl, cov.space)

```

Note that the rows of `cov.space` correspond to the columns of `params` and therefore the elements of `csl`:

```

> head(cov.space)
  Baseline Severity Rate of Decline Sex Carrier Status
1                9                0  F  NON-CARRIER
2               10                0  F  NON-CARRIER
3               11                0  F  NON-CARRIER
4               12                0  F  NON-CARRIER
5               13                0  F  NON-CARRIER
6               14                0  F  NON-CARRIER

```

In order to set up the calculator app, the user must then move the produced file `credsubs-shiny.RData` to the same directory as `server.R` and `ui.R` if it is not there already. The app server will then load the data file automatically when it is run.

5.5.2 A parametric example

We present also a more involved, parametric model fit using the package `nimble`. This example exhibits differences between `design` and `cov.space`, the default linear behavior when `design` is provided as an argument, and usage of `FUN`.

In this case, the data must be reparameterized to be entirely numeric:

```
library(nimble)

data <- read.csv("data-simplified.csv")

# Reparameterize factors so that reference=0, other=1
data$Treatment      <- as.numeric(relevel(data$Treatment,
                                           ref="Placebo")) - 1
data$Sex            <- as.numeric(relevel(data$Sex,
                                           ref="F")) - 1
data$Carrier.Status <- as.numeric(relevel(data$Carrier.Status,
                                           ref="NON-CARRIER")) - 1

data$Baseline.Severity <- scale(data$Baseline.Severity)
data$Rate.of.Decline  <- scale(data$Rate.of.Decline)

set.seed(1)
```

Next we must define the NIMBLE model, set constants, data, and initial values, compile, and run the sampler for 110,000 iterations, the first 10,000 of which will later be ignored as burn-in.

```
nimble.code <- nimbleCode({
  ### Likelihood ##
  for (i in 1:N) {
    # Linear predictor, expanded for clarity
    mean[i] <- beta[1] +
      beta[2] * Baseline.Severity[i] + beta[3] * Rate.of.Decline[i] +
      beta[4] * Sex[i] + beta[5] * Carrier.Status[i] +
      beta[6] * Sex[i] * Carrier.Status[i] +
      gamma[1] * Treatment[i] +
      gamma[2] * Baseline.Severity[i] * Treatment[i] +
      gamma[3] * Rate.of.Decline[i] * Treatment[i] +
      gamma[4] * Sex[i] * Treatment[i] +
      gamma[5] * Carrier.Status[i] * Treatment[i] +
      gamma[6] * Sex[i] * Carrier.Status[i] * Treatment[i]

    # Outcome
    Improvement[i] ~ dnorm(mean[i], tau) # tau = 1 / sigma ^ 2
  }

  # Vague priors for prognostic and baseline treatment effects
  # Conservative priors for treatment-covariate interactions
  for (j in 1:6) {
    beta[j] ~ dnorm(0, 10E-4)
    gamma[j] ~ dnorm(0, ifelse(j == 1, 10E-4, 1))
  }

  # Vague prior for error variance
  tau ~ dgamma(10E-4, 10E-4)
})
```

```

nimble.constants <- list(N=nrow(data))

nimble.data <- list(Treatment=as.numeric(data$Treatment),
                   Baseline.Severity=as.vector(data$Baseline.Severity),
                   Rate.of.Debate=as.vector(data$Rate.of.Debate),
                   Sex=as.numeric(data$Sex),
                   Carrier.Status=as.numeric(data$Carrier.Status),
                   Improvement=as.vector(data$Improvement))

nimble.inits <- list(beta=rep(0, 6),
                    gamma=rep(0, 6),
                    tau = 1)

nimble.model <- nimbleModel(code=nimble.code,
                           name = 'example',
                           constants = nimble.constants,
                           data = nimble.data,
                           inits = nimble.inits)

nimble.spec <- configureMCMC(nimble.model)
nimble.mcmc <- buildMCMC(nimble.spec)
C.model <- compileNimble(nimble.model)
C.mcmc <- compileNimble(nimble.mcmc, project = nimble.model)
C.mcmc$run(110000)
mcmc.trace <- as.matrix(C.mcmc$mvSamples)

```

Note that `mcmc.trace` is the trace of all of the parameters, of which we only need a few:

```

> colnames(mcmc.trace)
[1] "beta[1]" "beta[2]" "beta[3]" "beta[4]" "beta[5]"
[6] "beta[6]" "gamma[1]" "gamma[2]" "gamma[3]" "gamma[4]"
[11] "gamma[5]" "gamma[6]" "tau"

```

Thus we will keep only the interaction parameters (`gamma`) and construct `design` so that the PTE sample for each covariate point is the matrix product of `params` and a transposed row of `design`.

```
# Discard burn-in
keep <- 10001:110000

# Only some parameters needed for personalized treatment effect
pte.params <- mcmc.trace[keep, c("gamma[1]", "gamma[2]", "gamma[3]",
                                "gamma[4]", "gamma[5]", "gamma[6]")]

# Create data frame describing covariate space
design <- expand.grid(Treatment=1,
                    Baseline.Severity=9:49,
                    Rate.of.Decline=0:8,
                    Sex=c(0, 1),
                    Carrier.Status=c(0, 1))

# Since the model input was scaled, the covariate space must also be
design$Baseline.Severity <- scale(design$Baseline.Severity,
                                center=attr(data$Baseline.Severity,
                                              "scaled:center"),
                                scale=attr(data$Baseline.Severity,
                                             "scaled:scale"))
design$Rate.of.Decline <- scale(design$Rate.of.Decline,
                                center=attr(data$Rate.of.Decline,
                                              "scaled:center"),
                                scale=attr(data$Rate.of.Decline,
                                             "scaled:scale"))

# Add Sex-by-Carrier interaction
design$Sex.by.Car <- design$Sex * design$Carrier.Status
```

The `design` argument is then included in the call of `credsubs.level`:

```
cs1 <- credsubs.level(params=pte.params, design=design)
```

If we wish to define benefit as an effect greater than one conditional standard deviation of the outcome, we can append the error standard deviation draws ($\sigma = 1/\sqrt{\tau}$) to `pte.params`, instruct `FUN` to divide the previous matrix product by σ , and set `threshold` to 1:

```
pte.params.sd <- cbind(pte.params, 1 / sqrt(mcmc.trace[keep, "tau"]))
FUN <- function(x, params) {
  sd.col <- ncol(params)
  params[, -sd.col] %*% t(x) / params[, sd.col]
}
cs1.sd <- credsubs.level(params=pte.params.sd, design=design,
                        FUN=FUN, threshold=1)
```

In preparing the `shiny` calculator, the same `cov.space` object as in the previous example may be used here. Compare to the `design` object, in which continuous covariates have been standardized, factors converted to indicators, and extraneous columns (`Treatment`, `Sex.by.Car`) are present:

```
> head(design)
  Treatment Baseline.Severity Rate.of.Decline Sex Carrier.Status Sex.by.Car
1         1         -1.613593         -1.283438  0             0             0
2         1         -1.525112         -1.283438  0             0             0
3         1         -1.436632         -1.283438  0             0             0
4         1         -1.348151         -1.283438  0             0             0
5         1         -1.259670         -1.283438  0             0             0
6         1         -1.171190         -1.283438  0             0             0
```

Chapter 6

Conclusion

6.1 Summary of developments

In this thesis we have developed an approach identifying subgroups that benefit from treatment, which we believe can be used as a flexible confirmatory analysis. Specifically, the credible subgroups approach does not require pre-specification of subgroups, yet allows inferences that are not at inflated risk of producing false positive identifications. Thus a single trial may be used to both identify types of patients who benefit from treatment, and test the the personalized treatment effect throughout that benefiting subgroup with a controlled familywise Type I error rate.

Our approach, as first presented in Chapter 2, uses a Bayesian linear regression model to efficiently share information across the covariate space in an easily understandable way, and constructs simultaneous credible bands which take advantage of the induced dependencies among the posteriors of the PTEs to reduce the loss of power due to multiple testing adjustments. Furthermore, we presented an argument for interpreting credible subgroups from a frequentist prospective, and provided a sequential testing procedure for increased power. Chapter 3 generalized the method to arbitrary regression models, requiring at most a sample from the posterior distribution of the PTEs. Examples of use with generalized linear models, variable selection models, and semiparametric and nonparametric regression models are given.

Chapter 4 developed a framework for identifying benefiting subgroups in the presence of multiple endpoints and many treatments using adaptations of the decision-theoretic

concept of admissibility, which allows “benefit” to be defined without explicitly quantifying trade-offs via utility functions. While the discussion does not apply exclusively to credible subgroups, the credible subgroups method was the primary focus of the developments.

Chapter 5 gave details on implementation, including power calculations, standard error approximations, and diagnostics. Additionally, the chapter illustrated how software (provided in Appendix C) could be used to construct credible subgroups from the output of standard regression packages, and discussed possibilities for presenting the results of credible subgroups analyses, as well as using those results to implement treatment labeling and recommendations.

6.2 Significance of the work

A method that combines the steps of exploratory and confirmatory analysis is significant by virtue of negating the usual need for at least two separate trials to reliably identify and test a benefiting subgroup. However, this work also touches on and has the potential to influence wider areas: formalizing assumptions about treatment effect homogeneity or heterogeneity, evaluation of external validity, and general consideration of multiple testing in regression settings.

Initially, our aim was to discard the assumption of treatment effect homogeneity and begin identifying benefiting subpopulations from scratch. However, the choices of regression model (e.g., linear) and priors (how much shrinkage) constitute assumptions about the nature of heterogeneity in the treatment effect, though these assumptions are generally less stringent than that of total homogeneity. These mechanisms provide an opportunity to inject assumptions into analyses aiming to identify the benefiting population, but require that such assumptions be somewhat explicit: investigators must state that the regression model is linear or that the treatment-covariate interaction prior is tightly concentrated around zero, rather than rely on a shared implicit assumption of homogeneity. In fact, the assumption of treatment effect homogeneity is mathematically equivalent to specifying a linear interaction model with the priors for all interaction terms set to point masses at the origin, clearly a very strong assumption when stated explicitly. While the distinction between the two assumptions is primarily psychological,

we believe it is important to keep the consideration of effect heterogeneity at the front of investigators' minds.

Related to the formalization of assumptions is partial evaluation of external validity, the validity of causal inferences made from one study to non-study populations and situations. In context: does the treatment effect observed in the study hold outside of the study? Moving from an overall treatment effect to personalized treatment effects allows some treatment of external validity, at least with respect to variables included as predictive covariates in the study, and the assumptions made to share information across the predictive covariate space could arguably be used to make statements across a wider range of covariate values than that observed within the study. However, the strength of the credible subgroups with respect to this problem is negative rather than positive: the assumptions needed to generate conclusive inferences significantly outside of the observed covariate range (or even the observation-dense region of it) are likely immediately dubious. Thus investigators operating within the credible subgroups framework would be quantitatively compelled to include sufficiently diverse patients in the study in order to make convincing inferences about their target population. Of course, we must not mistake diversity with respect to observed predictive covariates for balance with respect to unobserved confounders that may otherwise jeopardize the external validity of a study.

This work does not exclusively increase the burden of evidence for proving benefit in populations, but also presents new opportunities to make the regulatory process more efficient, a primary aim of the 21st Century Cures Act. The credible subgroups approach, when used by treatment developers for in-house Phase II trials, can aid in identifying target populations for confirmatory trials in a principled manner. Confidence in such knowledge ahead of larger Phase III trials might greatly increase the odds of success. Additionally, using the credible subgroups approach in Phase III pharmaceutical trials and using the exclusive credible subgroup to label the drug allows regulators to approve treatments for populations for which sufficient evidence of benefit has been accumulated, while reserving judgment on usage in other populations until further study has been conducted. With appropriate advances in the method, such decisions could potentially even be made when clear evidence of benefit in certain subgroups is available in an ongoing long-term trial. Without heterogeneity-aware methods of analysis, trials

of treatments that are effective in a subset of the population may fail to detect that effect because it is attenuated by lack of effect in the rest of the population.

Finally, although the development of credible subgroups methods is motivated by a recurring problem in clinical trials, the approach itself is not restricted to such settings. Even beyond non-randomized studies of treatment effects, the general technique of constructing inclusion-exclusion pairs of bounding sets to estimate a subset may find uses in widely varied fields. In neuroimaging, a frequent task is to identify which voxels in a brain scan are active. In more computational data science fields such as network analysis, identifying which edges or entries of a sparse graph or matrix are non-zero is a common problem. In big data problems with a large number of covariates, identification of important subsets of variables is a primary concern. Though many methods exist for addressing these situations and others like them, these problems may benefit from the general credible subgroups approach, especially when multiplicity control is a concern or tests are highly correlated in a manner adequately described by regression models.

6.3 Future work

Several possible avenues for future work are discussed in this section. The first two deal with improvements to the approach as currently used, while the last two discuss situations not previously considered.

6.3.1 A “ 2α conjecture”

The simulation studies presented in Table 2.1 and Table 3.2 indicate that even though the nominal credible level is 80%, the restricted-space credible subgroups construction provide approximately 90% coverage when the underlying regression model is correctly specified. To understand why, recall that the most basic level, a $1 - \alpha$ credible subgroup pair is constructed by inverting a $1 - \alpha$ simultaneous credible band. That is, the posterior probability (and frequency, by the asymptotic argument) with which the true value of at least one personalized treatment effect lies outside of the band is at most α . However, a true PTE value lying outside of the simultaneous credible band does not necessarily result in a coverage failure by the credible subgroup pair: an error is made if the lower credible bound is greater than δ and the true PTE is not greater than δ , or if the upper

bound is less than δ and the true PTE greater. If the true PTE value is near δ at some covariate point and the credible band is unbiased there, the probability of a credible subgroup coverage failure there is approximately half the probability of a credible band coverage failure at the same location.

Unfortunately, the solution is not as simple as constructing $1 - \alpha$ credible subgroup pairs from a $1 - 2\alpha$ simultaneous credible band, even when α is small. The reason is that there is generally some probability that the simultaneous credible band will miss the true PTE value at more than one covariate point. Suppose that we have N independent signed tests each with a magnitude error rate of α' and an independent sign error rate of $1/2$. A credible band miss occurs when there is a magnitude error, and a credible subgroup miss occurs when there is both a magnitude and sign error. Thus the probability of any credible band miss (a simultaneous credible band miss) is $1 - (1 - \alpha')^N$ and that of a credible subgroup miss is $1 - (1 - \alpha'/2)^N$. If we were to instead assume that the probability of a credible subgroup miss is half the probability of a simultaneous credible band miss, the ratio of the true credible subgroup miss rate over the assumed rate would be

$$2 \frac{1 - (1 - \alpha'/2)^N}{1 - (1 - \alpha')^N} \xrightarrow{N \rightarrow \infty} 2, \quad (6.1)$$

showing that in the limit as $N \rightarrow \infty$, the actual credible subgroup coverage rate approaches the nominal rate rather than $1 - \alpha/2$.

Despite the theoretical danger of using a $1 - 2\alpha$ simultaneous credible band to construct a $1 - \alpha$ credible subgroup pair, our “ 2α conjecture” is that in many situations encountered in practice, the approximation works well. In particular, we expect that this would be the case when α is small (e.g., ≤ 0.05) and the effective number of independent tests is also small (e.g., ≤ 10) due to strong dependence among tests induced by the assumed regression model. The exact cases in which this approximation is reasonable, as well as how useful approximate familywise Type I error “guarantees” are, may be illuminated by a combination of analytical study, simulations, and practical experience.

6.3.2 A bootstrap counterpart

The theoretical underpinnings of a bootstrap counterpart to credible subgroups are straightforward: if the sample from the joint posterior of the PTEs is replaced by a

bootstrap sample representing the sampling distribution of estimates of those PTEs, many of the same approaches may be valid in the frequentist sense in a way that relies on the asymptotic justification of the bootstrap method rather than the asymptotic equivalence of the joint posterior and joint sampling distributions presented in this thesis. It remains to be seen whether the different asymptotic justification yields reliable credible subgroups at smaller sample sizes or with different modeling requirements. Furthermore, it would likely be necessary to employ certain shortcuts to reduce the computational burden to acceptable levels.

6.3.3 Large numbers of binary covariates

Much of the advantage of the credible subgroups approach over previous approaches to subgroup identification relate to the handling of continuous covariates. In addition to not requiring discretization of such covariates, our methods leverage the strong dependence of nearby PTEs using regression assumptions such as parametric forms for their variation over the covariate space. On the other hand, situations in which there are large numbers of binary covariates may require an approach focusing more on variable selection than common regression. Although some variable selection was discussed in Section 3.2.2, we have not attempted analyses on massive scales. It remains to be seen if the methods already presented can adequately handle these cases, or if new modeling or multiple testing techniques need to be developed.

6.3.4 Sequential and adaptive trial designs

Except for a brief discussion of sample size estimation, this thesis has focused entirely on the post-hoc analysis of clinical trials. However, there is ample opportunity to design trials from the start to take advantage of the credible subgroups approach through Bayesian adaptive designs [58]. One possible design is a special implementation of group sequential designs [59], in which the accumulated data are analyzed at pre-specified interim analyses throughout the course of the trial, and the trial may be stopped early either for clear evidence of benefit or futility in continuing. Incorporating credible subgroups could allow the trial to be stopped for regions of the covariate space at interim analyses, if those regions are contained in either D or S^c . Variations on this

these include adaptive enrichment designs [60] in which recruitment quotas are altered over the course of the trial to more efficiently obtain information about regions where the distinction between benefit and no benefit has not yet been made, and adaptive randomization designs [61, 44] in which randomization ratios are varied according to the results of interim credible subgroup analyses. In all of these cases, maintaining Type I error control (especially familywise) likely becomes substantially more difficult than when either credible subgroups or sequential/adaptive designs are used alone.

References

- [1] Food and Drug Administration safety and innovation act (FDASIA). *United States Code*, 2012.
- [2] US Food and Drug Administration. FDA action plan to enhance the collection and availability of demographic subgroup data. *FDA Report*, 2014.
- [3] The White House, Office of the Press Secretary. FACT SHEET: President Obama’s precision medicine initiative. *Press Release*, 2015.
- [4] H.R. 6, 21st Century Cures Act. *United States Code*, 2015.
- [5] S.A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.
- [6] US Food and Drug Administration. Guidance for industry: E9 statistical principles for clinical trials. *FDA Report*, 1998.
- [7] S.J. Pocock, S.E. Assmann, L.E. Enos, and L.E. Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21(19):2917–2930, 2002.
- [8] S.J. Ruberg, L. Chen, and Y. Wang. The mean does not mean as much anymore: finding sub-groups for tailored therapeutics. *Clinical Trials*, 7(5):574–583, 2010.
- [9] D.O. Dixon and R. Simon. Bayesian subset analysis. *Biometrics*, 47(3):871–881, 1991.
- [10] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Boca Raton, FL. CRC Press, 1984.

- [11] H.A. Chipman, E.I. George, and R.E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.
- [12] H.A. Chipman, E.I. George, and R.E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.
- [13] B. Freidlin and R. Simon. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research*, 11(21):7872–7878, 2005.
- [14] B. Freidlin, W. Jiang, and R. Simon. The cross-validated adaptive signature design. *Clinical Cancer Research*, 16(2):691–698, 2010.
- [15] X. Su, C.L. Tsai, H. Wang, D.M. Nickerson, and B. Li. Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research*, 10(Feb):141–158, 2009.
- [16] J.C. Foster, J.M.G. Taylor, and S.J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011.
- [17] I. Lipkovich, A. Dmitrienko, J. Denne, and G. Enas. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30(21):2601–2621, 2011.
- [18] Y. Xu, L. Trippa, P. Müller, and Y. Ji. Subgroup-based adaptive (SUBA) designs for multi-arm biomarker trials. *Statistics in Biosciences*, 8(1):1–22, 2016.
- [19] D. Altman, F. Song, C. Sakarovitch, J. Deeks, R. D’Amico, M. Bradburn, and A. Glenny A. Eastwood. Indirect comparisons of competing interventions. *Health Technology Assessment*, 2005.
- [20] H. Hong, H. Fu, K.L. Price, and B.P. Carlin. Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment. *Statistics in Medicine*, 34(20):2794–2819, 2015.
- [21] A. Burns and S. Iliffe. Alzheimer’s disease. *British Medical Journal*, 338:b158, 2009.

- [22] US Food and Drug Administration. FDA approves expanded use of treatment for patients with severe Alzheimer's disease. *Press Release*, 2006.
- [23] A. Qaseem, V. Snow, J.T. Cross, M.A. Forcica, R. Hopkins, P. Shekelle, A. Adelman, D. Mehr, K. Schellhase, D. Campos-Outcalt, P. Santaguida, D.K. Owens, and the Joint American College of Physicians/American Family Physicians Panel on Dementia. Current pharmacologic treatment of dementia: a clinical practice guideline from the American College of Physicians and the American Academy of Family Physicians. *Annals of Internal Medicine*, 148(5):370–378, 2008.
- [24] W.G. Rosen, R.C. Mohs, and K.L. Davis. A new rating scale for Alzheimer's disease. *The American Journal of Psychiatry*, 141(11):1356–1364, 1984.
- [25] D.V. Lindley and A.F.M. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(1):1–41, 1972.
- [26] H. Scheffé. *The Analysis of Variance*. New York. Wiley, 1959.
- [27] E. Uusipaikka. Exact confidence bands for linear regression over intervals. *Journal of the American Statistical Association*, 78(383):638–644, 1983.
- [28] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Boca Raton, FL. CRC Press, 2nd edition, 2004.
- [29] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- [30] S.P. Wright. Adjusted p-values for simultaneous inference. *Biometrics*, 48:1005–13, 1992.
- [31] R. Marcus, E. Peritz, and K.G. Ruben. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.
- [32] K.E. Peace and D.G.D. Chen. *Clinical Trial Methodology*. Boca Raton, FL. CRC Press, 2010.
- [33] M. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8:229–231, 1959.

- [34] R.B. Nelsen. *An Introduction to Copulas*. New York. Springer Science & Business Media, 2007.
- [35] NIMBLE Development Team. NIMBLE: An R package for programming with BUGS models, Version 0.4, 2015.
- [36] E.I. George and R.E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.
- [37] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 267–288, 1996.
- [38] T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.
- [39] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [40] J.L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.
- [41] J.O. Berger, X. Wang, and L. Shen. A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics*, 24(1):110–129, 2014.
- [42] B.A. Coull, D. Ruppert, and M.P. Wand. Simple incorporation of interactions into additive models. *Biometrics*, 57(2):539–545, 2001.
- [43] P.F. Thall, H.G. Sung, and E.H. Estey. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the American Statistical Association*, 97:29–39, 2002.
- [44] P.F. Thall, C. Logothetis, L.C. Pagliaro, S. Wen, M.A. Brown, D. Williams, and R.E. Millikan. Adaptive therapy for androgen-independent prostate cancer: a randomized selection trial of four regimens. *Journal of the National Cancer Institute*, 99(21):1613–1622, 2007.
- [45] D. Almirall, D.J. Lizotte, and S.A. Murphy. Comment. *Journal of the American Statistical Association*, 107(498):509–512, 2012.

- [46] D.J. Lizotte, M. Bowling, and S.A. Murphy. Linear fitted-Q iteration with multiple reward functions. *Journal of Machine Learning Research*, 13(Nov):3253–3295, 2012.
- [47] E.B. Laber, D.J. Lizotte, and B. Ferguson. Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, 70(1):53–61, 2014.
- [48] D.J. Lizotte and E.B. Laber. Multi-objective Markov decision processes for data-driven decision support. *Journal of Machine Learning Research*, 17:1–28, 2016.
- [49] J.L. Hill and Y.S. Su. Assessing lack of common support in causal inference using Bayesian nonparametrics: implications for evaluating the effect of breastfeeding on children’s cognitive outcomes. *The Annals of Applied Statistics*, 7(3):1386–1420, 2013.
- [50] M.H. Quenouille. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 11(1):68–84, 1949.
- [51] M.H. Quenouille. Notes on bias in estimation. *Biometrika*, 43(3/4):353–360, 1956.
- [52] J.W. Tukey. Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics*, 29(2):614–614, 1958.
- [53] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Boca Raton, FL. CRC Press, 1993.
- [54] S.J. Sheather and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 53(3):683–690, 1991.
- [55] RStudio, Inc. Shiny: a web application framework for R, 2017.
- [56] P.W.F. Wilson, R.B. D’Agostino, D. Levy, A.M. Belanger, H. Silbershatz, and W.B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.
- [57] J. Genest, R. McPherson, J. Frohlich, T. Anderson, N. Campbell, A. Carpentier, P. Couture, R. Dufour, G. Fodor, G. A. Francis, S. Grover, M. Gupta, R.A. Hegele, D.C. Lau, L. Leiter, G.F. Lewis, E. Lonn, G.B. Mancini, D. Ng, G.J.

- Pearson, A. Sniderman, J.A. Stone, and E. Ur. 2009 Canadian Cardiovascular Society/Canadian guidelines for the diagnosis and treatment of dyslipidemia and prevention of cardiovascular disease in the adult—2009 recommendations. *Canadian Journal of Cardiology*, 25(10):567–579, 2009.
- [58] S.M. Berry, B.P. Carlin, J.J. Lee, and P. Müller. *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton, FL. CRC Press, 2010.
- [59] S.J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.
- [60] N. Simon and R. Simon. Adaptive enrichment designs for clinical trials. *Biostatistics*, 14(4):613–625, 2013.
- [61] W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Appendix A

Proofs of theorems

A.1 Theorems for simultaneous credible bands

Theorem 1. *For a possibly infinite-dimensional vector of random variables \mathbf{Y} and vector of corresponding strictly increasing functions \mathbf{g} indexed by $\omega \in \Omega$,*

$$\mathbb{P} \left[\exists \omega \in \Omega : Y_\omega > g_\omega^{-1} \{ F_{\sup_{\Omega} g_\omega(Y_\omega)}^{-1}(1 - \alpha) \} \right] = \alpha. \quad (\text{A.1})$$

Proof.

$$\begin{aligned} & \mathbb{P} \left[\exists \omega \in \Omega : Y_\omega > g_\omega^{-1} \{ F_{\sup_{\Omega} g_\omega(Y_\omega)}^{-1}(1 - \alpha) \} \right], \\ &= \mathbb{P} \left[\exists \omega \in \Omega : g_\omega(Y_\omega) > F_{\sup_{\Omega} g_\omega(Y_\omega)}^{-1}(1 - \alpha) \right], \\ &= \mathbb{P} \left[\sup_{\Omega} g_\omega(Y_\omega) > F_{\sup_{\Omega} g_\omega(Y_\omega)}^{-1}(1 - \alpha) \right], \\ &= 1 - \mathbb{P} \left[\sup_{\Omega} g_\omega(Y_\omega) \leq F_{\sup_{\Omega} g_\omega(Y_\omega)}^{-1}(1 - \alpha) \right], \\ &= 1 - F_{\sup_{\Omega} g(Y_\omega)} \left[F_{\sup_{\Omega} g(Y_\omega)}^{-1}(1 - \alpha) \right], \\ &= 1 - (1 - \alpha), \\ &= \alpha. \end{aligned}$$

□

Corollary 1. *Simultaneous credible bands for the $\Delta_k^{tc}(\mathbf{x})$ may be constructed as*

$$\Delta_k^{tc}(\mathbf{x}) \in \mathbb{E}[\Delta_k^{tc}(\mathbf{x})] \pm \sqrt{W_\alpha^* \text{Var}[\Delta_k^{tc}(\mathbf{x})]} \quad (\text{A.2})$$

where W_α^* is the $1 - \alpha$ quantile of the distribution of

$$W = \sup_{(\mathbf{x}, k, (t, c))} \frac{\{\Delta_k^{tc}(\mathbf{x}) - \bar{\Delta}_k^{tc}(\mathbf{x})\}^2}{\text{Var}[\Delta_k^{tc}(\mathbf{x})]}.$$

Proof. Let $\omega = (\mathbf{x}, k, (t, c))$, $Y_\omega = \{\Delta_k^{tc}(\mathbf{x}) - \mathbb{E}[\Delta_k^{tc}(\mathbf{x})]\}^2$, and $g_\omega(Y_\omega) = Y_\omega / \mathbb{E}[Y_\omega]$. By Theorem 1 we have

$$\mathbb{P} \left[\exists (\mathbf{x}, k, t, c) : \{\Delta_k^{tc}(\mathbf{x}) - \mathbb{E}[\Delta_k^{tc}(\mathbf{x})]\}^2 > W_\alpha^* \text{Var}[\Delta_k^{tc}(\mathbf{x})] \right] = \alpha.$$

The validity of (A.2) follows immediately. \square

Lemma 1. *If g in Theorem 1 is such that the $g_\omega(Y_\omega)$ identically distributed, then*

$$\mathbb{P} \left[Y_\omega > g_\omega^{-1} \left\{ F_{\sup_{\Omega} g_\omega(Y_\omega)}^{-1}(1 - \alpha) \right\} \right] \quad (\text{A.3})$$

is constant over Ω .

Proof. Apply g_ω to both sides of the relation. \square

Theorem 2. *If $g_\omega = F_{Y_\omega}$ for all $\omega \in \Omega$ then*

$$\mathbb{P} \left[Y_\omega > g_\omega^{-1} \left\{ F_{\sup_{\Omega} g_\omega(Y_\omega)}^{-1}(1 - \alpha) \right\} \right] \quad (\text{A.4})$$

is constant over Ω .

Proof. If $g_\omega = F_{Y_\omega}$ for all $\omega \in \Omega$ then $g_\omega(Y_\omega) \sim \text{Uniform}(0, 1)$ for all $\omega \in \Omega$, and Lemma 1 yields the result. \square

Theorem 3. Suppose \mathbf{Y} is a possibly infinite-dimensional vector of random variables indexed by $\omega \in \Omega$. Let $F_{Y_\omega}(y) = \mathbb{P}[Y_\omega \leq y]$, $G_{Y_\omega}(y) = \mathbb{P}[Y_\omega < y]$, $F_{Y_\omega}^{-1}(p) = \inf \{y \in \mathbb{R} : p \leq F_{Y_\omega}(y)\}$, and $G_{Y_\omega}^{-1}(p) = \sup \{y \in \mathbb{R} : p \geq G_{Y_\omega}(y)\}$. If W_α^* is the $1 - \alpha$ quantile of the distribution of

$$W = \sup_{\omega \in \Omega} \max \{1 - F_{Y_\omega}(Y_\omega), G_{Y_\omega}(Y_\omega)\}, \quad (\text{A.5})$$

then

$$\mathbb{P} [\forall \omega \in \Omega, F_{Y_\omega}^{-1}(1 - W_\alpha^*) \leq Y_\omega \leq G_{Y_\omega}^{-1}(W_\alpha^*)] \geq 1 - \alpha. \quad (\text{A.6})$$

Proof.

$$\begin{aligned} & \mathbb{P} [\forall \omega \in \Omega, F_{Y_\omega}^{-1}(1 - W_\alpha^*) \leq Y_\omega \leq G_{Y_\omega}^{-1}(W_\alpha^*)], \\ &= \mathbb{P} \left[\sup_{\omega \in \Omega} \max \{F_{Y_\omega}^{-1}(1 - W_\alpha^*) - Y_\omega, Y_\omega - G_{Y_\omega}^{-1}(W_\alpha^*)\} \leq 0 \right], \\ &= \mathbb{P} \left[\sup_{\omega \in \Omega} \max \{F_{Y_\omega} [F_{Y_\omega}^{-1}(1 - W_\alpha^*)] - F_{Y_\omega}(Y_\omega), G_{Y_\omega}(Y_\omega) - G_{Y_\omega} [G_{Y_\omega}^{-1}(W_\alpha^*)]\} \leq 0 \right], \\ &= \mathbb{P} \left[\sup_{\omega \in \Omega} \max \{1 - W_\alpha^* - F_{Y_\omega}(Y_\omega), G_{Y_\omega}(Y_\omega) - W_\alpha^*\} \leq 0 \right], \\ &= \mathbb{P} \left[\sup_{\omega \in \Omega} \max \{1 - F_{Y_\omega}(Y_\omega), G_{Y_\omega}(Y_\omega)\} \leq W_\alpha^* \right], \\ &= \mathbb{P} [W \leq W_\alpha^*], \\ &= 1 - \alpha. \end{aligned}$$

□

A.2 Theorems for sequential testing procedures

Theorem 4 (Batch Step-Down Testing Procedure). *Let C be the restricted covariate space, $\boldsymbol{\theta}$ be the vector of all model parameters, and $H_{\mathbf{z}} = \{\boldsymbol{\theta} : \Delta(\mathbf{z}) = \delta\}$. The following testing procedure controls the familywise Type I error rate at level α .*

-
-
- 1 Let $M = 1$, $T_0 = C$, and $R_0 = \emptyset$ be the starting iteration, base test set, and base rejection set, respectively;
 - 2 **repeat**
 - 3 Let $T_M = T_{M-1} \setminus R_{M-1} = C \setminus (\bigcup_{m < M} R_m)$ be the new test set;
 - 4 Construct the two-sided $1 - \alpha$ restricted covariate space simultaneous confidence band (2.10) for $\Delta(\mathbf{z})$ over all $\mathbf{z} \in T_M$;
 - 5 Let R_M be the set of \mathbf{z} for which the band does not contain zero;
 - 6 Increment M ;
 - 7 **until** $R_M = \emptyset$;
 - 8 Reject $H_{\mathbf{z}}$ for all $\mathbf{z} \in \bigcup_{m < M} R_m$.
-

The proof of Theorem 4 relies on showing that the batch step-down testing procedure is a *closed testing procedure* [31]. Let \mathcal{H} be a collection of hypotheses closed under intersection: $H_{\mathbf{z}}, H_{\boldsymbol{\zeta}} \in \mathcal{H}$ implies $H_{\mathbf{z}} \cap H_{\boldsymbol{\zeta}} \in \mathcal{H}$. Furthermore, let $\phi_{\mathbf{z}}$ be an α -level test of $H_{\mathbf{z}}$ so that $\phi_{\mathbf{z}} = 1$ if and only if it rejects $H_{\mathbf{z}}$ locally (independent of other hypotheses and tests). Then a *closed testing procedure* is a procedure which rejects $H_{\mathbf{z}}$ if and only if $\phi_{\boldsymbol{\zeta}} = 1$ for all $\boldsymbol{\zeta}$ such that $H_{\boldsymbol{\zeta}} \subseteq H_{\mathbf{z}} \in \mathcal{H}$. Any closed testing procedure controls the family-wise Type I error rate at level α because in order to reject at least one true hypothesis the procedure must reject the intersection of all true hypotheses, which is tested by an α -level test.

Proof of Theorem 4. Let $H_U = \bigcap_{\mathbf{z} \in U} H_{\mathbf{z}}$ be the hypothesis that $\Delta(\mathbf{z}) = \delta$ for all $\mathbf{z} \in U$, and ϕ_U be the local test of that hypothesis which rejects if and only if exists an $\mathbf{z} \in U$ for which the band does not include δ . We first show that if $\phi_U = 1$ and the band over

U did not include δ at $\mathbf{z} \in V \subset U$, then $\phi_V = 1$. Suppose $\phi_U = 1$ and the band over U did not contain δ at $\mathbf{z} \in V \subset U$. Let $W_{\alpha,U}^*$ be the $1 - \alpha$ quantile of the distribution of

$$W_U = \sup_{\mathbf{z} \in U} \frac{\|\bar{\Delta}(\mathbf{z}) - \Delta(\mathbf{z})\|}{\sqrt{\text{Var}[\bar{\Delta}(\mathbf{z})]}}.$$

Note that since $V \subset U$, we have $W_V \leq W_U$, and thus $W_{\alpha,V}^* \leq W_{\alpha,U}^*$. Then the band over V is nowhere wider than the band over U, and since the band over U did not contain zero at \mathbf{z} , the band over V also does not contain zero at \mathbf{z} , so $\phi_V = 1$.

We now show that when a point \mathbf{z} is marked for rejection, it is rejected by all intersections in \mathcal{H} involving $H_{\mathbf{z}}$. Let $\mathbf{z} \in R_1$, i.e., a point whose null-effect hypothesis was marked for rejection on the first iteration. Since $H_{\mathbf{z}}$ was marked on the first iteration, there is at least one \mathbf{z} (itself) in C for which the band does not include zero; thus $\phi_C = 1$. Additionally, any other hypothesis in \mathcal{H} which is an intersection involving $H_{\mathbf{z}}$ is a hypothesis H_V with $\mathbf{z} \in V \subset C$, and thus is also rejected by the corresponding local test. Thus $H_{\mathbf{z}}$ may be globally rejected.

Consider now $\mathbf{z} \in R_M$, $M > 1$. Every $\zeta \in C \setminus T_M$ has previously had its hypothesis globally rejected, thus any local test of a hypothesis for a set containing that point has already been locally rejected. Therefore we need only consider hypotheses H_U such that $U \subseteq T_M$. The argument for points in R_1 may then be reused, replacing C with T_M . Thus the procedure is a closed testing procedure, and therefore controls the family-wise Type I error rate at α . \square

Remark 1. *Theorem 4 applies to posterior credible bands insofar as they correspond to the confidence band via the asymptotic joint normality of the posterior of $\Delta(\mathbf{z})$.*

Appendix B

Expanded simulation results

B.1 Additional simulations for basic credible subgroups

B.2 Comparison of parametric, semiparametric, and non-parametric regression

Truth	Method	Total Coverage	D Coverage	S Coverage	Pair Size	Heterog. Tests
$\gamma = (0, 0, 0)$	PB	0.46	0.46	1.00	0.75	0.18
	RCS	0.88	0.88	1.00	0.95	0.18
	HPD	0.91	0.91	1.00	0.97	0.18
	PW	0.43	0.43	1.00	0.59	0.18
$\gamma = (0, 0, 1)$	PB	0.82	0.87	0.94	0.25	1.00
	RCS	0.94	0.95	0.99	0.34	1.00
	HPD	0.96	0.97	0.99	0.38	1.00
	PW	0.46	0.67	0.77	0.13	1.00
$\gamma = (0, 1, 0)$	PB	0.55	0.55	0.99	0.55	0.45
	RCS	0.87	0.87	1.00	0.78	0.45
	HPD	0.91	0.91	1.00	0.82	0.45
	PW	0.47	0.47	0.97	0.39	0.45
$\gamma = (0, 1, 1)$	PB	0.77	0.84	0.93	0.25	1.00
	RCS	0.92	0.93	0.98	0.35	1.00
	HPD	0.95	0.96	0.99	0.38	1.00
	PW	0.41	0.61	0.76	0.14	1.00
$\gamma = (1, 0, 0)$	PB	0.99	1.00	0.99	0.25	0.18
	RCS	1.00	1.00	1.00	0.50	0.18
	HPD	1.00	1.00	1.00	0.56	0.18
	PW	0.97	1.00	0.97	0.13	0.18
$\gamma = (1, 1, 1)$	PB	0.73	0.79	0.94	0.24	1.00
	RCS	0.92	0.93	0.99	0.33	1.00
	HPD	0.94	0.95	1.00	0.35	1.00
	PW	0.43	0.62	0.80	0.15	1.00
Near-Linear	PB	0.64	0.86	0.77	0.62	0.56
	RCS	0.92	0.97	0.95	0.84	0.56
	HPD	0.94	0.98	0.97	0.87	0.56
	PW	0.38	0.74	0.61	0.42	0.56
Threshold	PB	0.76	0.85	0.90	0.44	0.92
	RCS	0.93	0.95	0.98	0.61	0.92
	HPD	0.95	0.96	0.99	0.65	0.92
	PW	0.42	0.65	0.74	0.24	0.92
Non-Monotone	PB	0.20	0.41	0.70	0.73	0.18
	RCS	0.80	0.82	0.96	0.93	0.18
	HPD	0.85	0.86	0.98	0.95	0.18
	PW	0.16	0.30	0.60	0.56	0.18

Table B.1: Coverage and model fit statistics for 80% credible subgroup pairs (n=40).

Truth	Method	Total Coverage	D Coverage	S Coverage	Pair Size	Heterog. Tests
$\gamma = (0, 0, 0)$	PB	0.45	0.45	1.00	0.75	0.18
	RCS	0.88	0.88	1.00	0.95	0.18
	HPD	0.92	0.92	1.00	0.97	0.18
	PW	0.43	0.43	1.00	0.59	0.18
$\gamma = (0, 0, 1)$	PB	0.79	0.85	0.93	0.14	1.00
	RCS	0.94	0.95	0.99	0.19	1.00
	HPD	0.96	0.97	0.99	0.21	1.00
	PW	0.45	0.64	0.77	0.08	1.00
$\gamma = (0, 1, 0)$	PB	0.53	0.53	1.00	0.41	0.77
	RCS	0.89	0.89	1.00	0.58	0.77
	HPD	0.92	0.92	1.00	0.62	0.77
	PW	0.53	0.53	1.00	0.31	0.77
$\gamma = (0, 1, 1)$	PB	0.76	0.83	0.92	0.14	1.00
	RCS	0.92	0.94	0.98	0.20	1.00
	HPD	0.95	0.96	0.99	0.22	1.00
	PW	0.45	0.63	0.76	0.08	1.00
$\gamma = (1, 0, 0)$	PB	1.00	1.00	1.00	0.04	0.18
	RCS	1.00	1.00	1.00	0.17	0.18
	HPD	1.00	1.00	1.00	0.21	0.18
	PW	1.00	1.00	1.00	0.02	0.18
$\gamma = (1, 1, 1)$	PB	0.76	0.84	0.92	0.16	1.00
	RCS	0.93	0.95	0.98	0.22	1.00
	HPD	0.96	0.96	0.99	0.24	1.00
	PW	0.44	0.64	0.77	0.09	1.00
Near-Linear	PB	0.68	0.92	0.76	0.50	0.89
	RCS	0.90	0.98	0.92	0.68	0.89
	HPD	0.93	0.99	0.94	0.72	0.89
	PW	0.37	0.77	0.55	0.29	0.89
Threshold	PB	0.78	0.87	0.91	0.26	1.00
	RCS	0.93	0.95	0.98	0.35	1.00
	HPD	0.96	0.97	0.99	0.39	1.00
	PW	0.42	0.65	0.72	0.14	1.00
Non-Monotone	PB	0.20	0.34	0.77	0.70	0.18
	RCS	0.77	0.79	0.97	0.92	0.18
	HPD	0.83	0.84	0.98	0.94	0.18
	PW	0.14	0.24	0.68	0.53	0.18

Table B.2: Coverage and model fit statistics for 80% credible subgroup pairs (n=100).

Truth	Method	Total Coverage	D Coverage	S Coverage	Pair Size	Heterog. Tests
$\gamma = (0, 0, 0)$	PB	0.43	0.43	1.00	0.76	0.21
	RCS	0.88	0.88	1.00	0.95	0.21
	HPD	0.91	0.91	1.00	0.97	0.21
	PW	0.41	0.41	1.00	0.59	0.21
$\gamma = (0, 0, 1)$	PB	0.79	0.84	0.95	0.07	1.00
	RCS	0.96	0.96	0.99	0.10	1.00
	HPD	0.97	0.97	0.99	0.11	1.00
	PW	0.56	0.66	0.87	0.04	1.00
$\gamma = (0, 1, 0)$	PB	0.48	0.48	1.00	0.34	1.00
	RCS	0.92	0.92	1.00	0.48	1.00
	HPD	0.93	0.93	1.00	0.48	1.00
	PW	0.54	0.54	1.00	0.29	1.00
$\gamma = (0, 1, 1)$	PB	0.77	0.82	0.95	0.07	1.00
	RCS	0.95	0.96	0.99	0.10	1.00
	HPD	0.96	0.97	0.99	0.11	1.00
	PW	0.54	0.64	0.87	0.04	1.00
$\gamma = (1, 0, 0)$	PB	1.00	1.00	1.00	0.00	0.21
	RCS	1.00	1.00	1.00	0.00	0.21
	HPD	1.00	1.00	1.00	0.00	0.21
	PW	1.00	1.00	1.00	0.00	0.21
$\gamma = (1, 1, 1)$	PB	0.77	0.82	0.94	0.08	1.00
	RCS	0.93	0.94	0.99	0.12	1.00
	HPD	0.96	0.96	1.00	0.13	1.00
	PW	0.52	0.65	0.87	0.05	1.00
Near-Linear	PB	0.72	0.97	0.74	0.26	1.00
	RCS	0.87	0.99	0.87	0.36	1.00
	HPD	0.92	1.00	0.92	0.39	1.00
	PW	0.37	0.88	0.43	0.14	1.00
Threshold	PB	0.79	0.86	0.93	0.12	1.00
	RCS	0.96	0.96	0.99	0.17	1.00
	HPD	0.97	0.98	0.99	0.18	1.00
	PW	0.50	0.66	0.80	0.07	1.00
Non-Monotone	PB	0.15	0.18	0.90	0.61	0.22
	RCS	0.61	0.62	0.98	0.86	0.22
	HPD	0.70	0.70	0.99	0.89	0.22
	PW	0.08	0.10	0.82	0.43	0.22

Table B.3: Coverage and model fit statistics for 80% credible subgroup pairs (n=350).

Truth	Method	Sensitivity of D	Specificity of D	Sensitivity of S	Specificity of S
$\gamma = (0, 0, 0)$	PB	–	0.87	–	0.12
	RCS	–	0.97	–	0.02
	HPD	–	0.98	–	0.01
	PW	–	0.79	–	0.20
$\gamma = (0, 0, 1)$	PB	0.76	0.99	1.00	0.74
	RCS	0.67	1.00	1.00	0.64
	HPD	0.64	1.00	1.00	0.61
	PW	0.87	0.98	0.99	0.84
$\gamma = (0, 1, 0)$	PB	0.68	0.83	1.00	0.05
	RCS	0.38	0.95	1.00	0.01
	HPD	0.33	0.96	1.00	0.00
	PW	0.79	0.71	0.99	0.13
$\gamma = (0, 1, 1)$	PB	0.81	0.99	1.00	0.64
	RCS	0.75	1.00	1.00	0.52
	HPD	0.72	1.00	1.00	0.48
	PW	0.89	0.97	0.99	0.78
$\gamma = (1, 0, 0)$	PB	0.75	–	1.00	–
	RCS	0.50	–	1.00	–
	HPD	0.44	–	1.00	–
	PW	0.87	–	1.00	–
$\gamma = (1, 1, 1)$	PB	0.87	0.97	1.00	0.39
	RCS	0.82	0.99	1.00	0.25
	HPD	0.80	0.99	1.00	0.21
	PW	0.92	0.93	0.99	0.59
Near-Linear	PB	0.28	0.98	0.96	0.41
	RCS	0.13	1.00	1.00	0.19
	HPD	0.10	1.00	1.00	0.15
	PW	0.45	0.95	0.92	0.58
Threshold	PB	0.56	0.99	0.99	0.54
	RCS	0.40	1.00	1.00	0.38
	HPD	0.35	1.00	1.00	0.34
	PW	0.74	0.97	0.98	0.72
Non-Monotone	PB	0.21	0.81	0.94	0.07
	RCS	0.06	0.95	0.99	0.01
	HPD	0.04	0.96	1.00	0.01
	PW	0.33	0.71	0.89	0.14

Table B.4: Diagnostic properties of 80% credible subgroup pairs (n=40).

Truth	Method	Sensitivity of D	Specificity of D	Sensitivity of S	Specificity of S
$\gamma = (0, 0, 0)$	PB	–	0.88	–	0.13
	RCS	–	0.98	–	0.02
	HPD	–	0.98	–	0.02
	PW	–	0.80	–	0.21
$\gamma = (0, 0, 1)$	PB	0.87	1.00	1.00	0.85
	RCS	0.82	1.00	1.00	0.79
	HPD	0.81	1.00	1.00	0.78
	PW	0.92	0.99	0.99	0.90
$\gamma = (0, 1, 0)$	PB	0.93	0.83	1.00	0.08
	RCS	0.78	0.96	1.00	0.01
	HPD	0.73	0.97	1.00	0.01
	PW	0.96	0.73	1.00	0.16
$\gamma = (0, 1, 1)$	PB	0.89	0.99	1.00	0.80
	RCS	0.85	1.00	1.00	0.73
	HPD	0.84	1.00	1.00	0.71
	PW	0.93	0.98	0.99	0.87
$\gamma = (1, 0, 0)$	PB	0.96	–	1.00	–
	RCS	0.83	–	1.00	–
	HPD	0.79	–	1.00	–
	PW	0.98	–	1.00	–
$\gamma = (1, 1, 1)$	PB	0.91	0.99	1.00	0.61
	RCS	0.88	1.00	1.00	0.49
	HPD	0.87	1.00	1.00	0.45
	PW	0.95	0.96	0.99	0.74
Near-Linear	PB	0.39	0.99	0.98	0.58
	RCS	0.24	1.00	0.99	0.40
	HPD	0.20	1.00	1.00	0.35
	PW	0.60	0.98	0.94	0.75
Threshold	PB	0.75	0.99	1.00	0.73
	RCS	0.66	1.00	1.00	0.64
	HPD	0.62	1.00	1.00	0.60
	PW	0.86	0.98	0.98	0.83
Non-Monotone	PB	0.26	0.79	0.95	0.06
	RCS	0.08	0.94	1.00	0.01
	HPD	0.06	0.96	1.00	0.00
	PW	0.39	0.67	0.91	0.11

Table B.5: Diagnostic properties of 80% credible subgroup pairs (n=100).

Truth	Method	Sensitivity of D	Specificity of D	Sensitivity of S	Specificity of S
$\gamma = (0, 0, 0)$	PB	–	0.88	–	0.12
	RCS	–	0.98	–	0.02
	HPD	–	0.98	–	0.02
	PW	–	0.80	–	0.21
$\gamma = (0, 0, 1)$	PB	0.94	1.00	1.00	0.92
	RCS	0.91	1.00	1.00	0.89
	HPD	0.91	1.00	1.00	0.88
	PW	0.96	0.99	1.00	0.94
$\gamma = (0, 1, 0)$	PB	1.00	0.81	1.00	0.12
	RCS	1.00	0.97	1.00	0.02
	HPD	1.00	0.98	1.00	0.01
	PW	1.00	0.76	1.00	0.17
$\gamma = (0, 1, 1)$	PB	0.95	1.00	1.00	0.90
	RCS	0.93	1.00	1.00	0.86
	HPD	0.92	1.00	1.00	0.85
	PW	0.97	0.99	1.00	0.93
$\gamma = (1, 0, 0)$	PB	1.00	–	1.00	–
	RCS	1.00	–	1.00	–
	HPD	1.00	–	1.00	–
	PW	1.00	–	1.00	–
$\gamma = (1, 1, 1)$	PB	0.96	0.99	1.00	0.80
	RCS	0.94	1.00	1.00	0.72
	HPD	0.93	1.00	1.00	0.70
	PW	0.97	0.98	1.00	0.86
Near-Linear	PB	0.63	1.00	0.99	0.83
	RCS	0.53	1.00	0.99	0.75
	HPD	0.49	1.00	1.00	0.72
	PW	0.76	1.00	0.96	0.91
Threshold	PB	0.89	1.00	1.00	0.86
	RCS	0.84	1.00	1.00	0.82
	HPD	0.83	1.00	1.00	0.81
	PW	0.93	0.99	0.99	0.91
Non-Monotone	PB	0.39	0.69	0.99	0.02
	RCS	0.15	0.89	1.00	0.00
	HPD	0.12	0.92	1.00	0.00
	PW	0.55	0.55	0.96	0.06

Table B.6: Diagnostic properties of 80% credible subgroup pairs (n=350).

Data Generating Mechanism	Model	Coverage	Sensitivity of D	Step-Down Efficiency
Null Effect	Linear	0.88	—	—
	Spline	0.93	—	—
	BART	1.00	—	—
Binary	Linear	0.92	0.51	1.04
	Spline	0.95	0.13	1.04
	BART	0.96	0.25	1.12
Linear	Linear	0.89	0.72	1.04
	Spline	0.97	0.32	1.07
	BART	1.00	0.12	1.08
Near-Linear	Linear	0.87	0.37	1.08
	Spline	0.98	0.07	1.10
	BART	0.98	0.01	1.06
Threshold	Linear	0.84	0.40	1.05
	Spline	0.97	0.11	1.04
	BART	0.97	0.06	1.12
Non-Monotone	Linear	0.60	0.07	1.10
	Spline	0.96	0.20	1.05
	BART	0.88	0.12	1.19

Table B.7: Simulation study results. Operating characteristics of 80% credible subgroups with $n = 25$ patients in each study arm. Struck-through sensitivities indicate insufficient coverage and should be treated with caution.

Data Generating Mechanism	Model	Coverage	Sensitivity of D	Step-Down Efficiency
Null Effect	Linear	0.86	—	—
	Spline	0.92	—	—
	BART	0.99	—	—
Binary	Linear	0.91	0.81	1.03
	Spline	0.95	0.28	1.04
	BART	0.96	0.52	1.08
Linear	Linear	0.87	0.84	1.03
	Spline	0.96	0.60	1.12
	BART	1.00	0.50	1.11
Near-Linear	Linear	0.81	0.58	1.06
	Spline	0.97	0.18	1.15
	BART	0.99	0.05	1.16
Threshold	Linear	0.76	0.66	1.04
	Spline	0.96	0.26	1.05
	BART	0.97	0.16	1.08
Non-Monotone	Linear	0.46	0.29	1.08
	Spline	0.96	0.40	1.06
	BART	0.91	0.21	1.12

Table B.8: Simulation study results. Operating characteristics of 80% credible subgroups with $n = 50$ patients in each study arm. Struck-through sensitivities indicate insufficient coverage and should be treated with caution.

Data Generating Mechanism	Model	Coverage	Sensitivity of D	Step-Down Efficiency
Null Effect	Linear	0.89	—	—
	Spline	0.92	—	—
	BART	1.00	—	—
Binary	Linear	0.90	0.92	1.02
	Spline	0.94	0.41	1.05
	BART	0.97	0.67	1.07
Linear	Linear	0.88	0.87	1.02
	Spline	0.97	0.71	1.10
	BART	1.00	0.65	1.08
Near-Linear	Linear	0.78	0.67	1.05
	Spline	0.98	0.29	1.18
	BART	1.00	0.16	1.22
Threshold	Linear	0.69	0.79	1.03
	Spline	0.98	0.39	1.07
	BART	0.99	0.33	1.07
Non-Monotone	Linear	0.32	0.41	1.09
	Spline	0.97	0.54	1.06
	BART	0.95	0.34	1.09

Table B.9: Simulation study results. Operating characteristics of 80% credible subgroups with $n = 75$ patients in each study arm. Struck-through sensitivities indicate insufficient coverage and should be treated with caution.

Appendix C

Software Code

C.1 credsubs.R

```
if (!require(ff)) {
  warning("Package 'ff' required to use function 'credsubs.level'",
         "with option z.store='disk'")
}

to.Fx <- function(x) {
  # A faster version of ecdf(x)(x)
  rank(x, ties.method="max")/length(x)
}

sim.cred.band <- function(params, design=NULL,
                          FUN=function(x, params) { params %*% t(x) },
                          est.fun=mean,
                          var.fun=sd,
                          return.w=TRUE,
                          cred.level=0.95,
                          method=c('asymptotic', 'quantile'),
                          verbose=TRUE) {
```

```

# Validate and shape input
params <- as.matrix(params)
M <- nrow(params)

if (is.null(design)) {
  N <- ncol(params)
  nonpar <- TRUE
} else {
  design <- data.matrix(design)
  N <- nrow(design)
  nonpar <- FALSE
}

method <- method[1]

sim.cred.band <- list(upper=rep(NA, N),
                    lower=rep(NA, N),
                    est=rep(NA, N),
                    est.fun=est.fun,
                    var=rep(NA, N),
                    var.fun=var.fun,
                    cred.level=cred.level,
                    method=method,
                    W.crit=NA)
class(sim.cred.band) <- 'sim.cred.band'

est <- var <- numeric(N)

if (method == 'asymptotic') {
  m <- numeric(N)
  s <- numeric(N)
} else if (method == 'quantile') {} else {
  warning("method must be one of 'asymptotic' or 'quantile'. Given: ",
         method)
  return(sim.cred.band)
}

```

```

# Compute W
W <- rep(-Inf, M)

# This is iterative to avoid storing an NxM matrix (or two)
for (i in 1:N) {

  if (verbose && (i %% 100 == 0)) {
    cat(i, "/", N, "\n")
  }

  if (nonpar) {
    fx <- params[, i]
  } else {
    fx <- FUN(design[i, , drop=FALSE], params)
  }

  est[i] <- est.fun(fx)
  var[i] <- var.fun(fx)

  if (method == 'asymptotic') {
    m[i] <- mean(fx)
    s[i] <- sd(fx)
    z <- abs(fx - m[i]) / s[i]
  } else {
    Fx <- to.Fx(fx)
    Gx <- 1 - to.Fx(-fx)
    z <- pmax(1-Fx, Gx)
  }
  W <- pmax(W, z)
}

if (return.w) {
  sim.cred.band$W <- W
}

```

```

# Estimate W.crit
W.crit <- quantile(W, cred.level, type=1)
sim.cred.band$W.crit <- W.crit

# Compute bounds
if (method == 'asymptotic') {
  sim.cred.band$upper <- m + W.crit * s
  sim.cred.band$lower <- m - W.crit * s
} else {
  if (nonpar) {
    bounds <- apply(params, 2, function(fx) {
      upper <- -quantile(-fx, prob=1-W.crit, type=1)
      lower <- quantile(fx, prob=1-W.crit, type=1)
      c(lower, upper)
    })
  } else {
    bounds <- apply(design, 1, function(x, params) {
      fx <- FUN(t(x), params)
      upper <- -quantile(-fx, prob=1-W.crit, type=1)
      lower <- quantile(fx, prob=1-W.crit, type=1)
      c(lower, upper)
    }, params=params)
  }

  sim.cred.band$lower <- bounds[1, ]
  sim.cred.band$upper <- bounds[2, ]
}

sim.cred.band$est <- est
sim.cred.band$var <- var

sim.cred.band
}

```

```

credsubs <- function(params, design=NULL,
                    FUN=function(x, params) { params %*% t(x) },
                    cred.level=0.95,
                    threshold=0,
                    method=c('asymptotic', 'quantile'),
                    step.down=TRUE,
                    verbose=TRUE) {

  # Validate and shape input
  params <- as.matrix(params)
  M <- nrow(params)

  if (is.null(design)) {
    N <- ncol(params)
    nonpar <- TRUE
  } else {
    design <- data.matrix(design)
    N <- nrow(design)
    nonpar <- FALSE
  }

  if (verbose) {
    cat("Computing credible subgroups over", N, "points using",
        M, "posterior draws.\n")
  }

  method <- method[1]

  credsubs <- list(exclusive=rep(NA, N),
                  inclusive=rep(NA, N),
                  cred.level=cred.level,
                  threshold=threshold,
                  method=method,
                  step.down=step.down,
                  W.crit=NA)
  class(credsubs) <- 'credsubs'
}

```

```

if (method == 'asymptotic') {
  m <- numeric(N)
  s <- numeric(N)
} else if (method == 'quantile') {} else {
  warning("method must be one of 'asymptotic' or 'quantile'. Given: ",
          method)
  return(credsubs)
}

credsubs$exclusive <- rep(FALSE, N)
credsubs$inclusive <- rep(TRUE, N)

test.set <- 1:N
reject.set <- numeric(0)

repeat {
  test.set <- setdiff(test.set, reject.set)
  if (nonpar) {
    test.par <- test.set
  } else {
    test.par <- 1:ncol(params)
  }

  sim.cred.band <- sim.cred.band(params=params[, test.par, drop=FALSE],
                                design=design[test.set, , drop=FALSE],
                                FUN=FUN,
                                cred.level=cred.level,
                                method=method,
                                verbose=verbose)

  over.set <- test.set[sim.cred.band$lower > threshold]
  under.set <- test.set[sim.cred.band$upper < threshold]
  credsubs$exclusive[over.set] <- TRUE
  credsubs$inclusive[under.set] <- FALSE
  reject.set <- union(over.set, under.set)
}

```



```

    if (verbose) {
      cat(length(test.set), "hypotheses tested,",
          length(reject.set), "rejected.\n")
    }
    if (!step.down ||
        length(reject.set) == 0 ||
        length(test.set) == length(reject.set)) {
      break
    }
  }
}

credsubs
}

credsubs.level <- function(params, design=NULL,
                           FUN=function(x, params) { params %*% t(x) },
                           threshold=0,
                           method=c('asymptotic', 'quantile'),
                           W.probs=NULL,
                           step.down=TRUE,
                           verbose=TRUE,
                           z.store=c("ram", "recompute", "disk")) {

  # Validate and shape input
  params <- as.matrix(params)
  M <- nrow(params)

  if (is.null(design)) {
    N <- ncol(params)
    nonpar <- TRUE
  } else {
    design <- data.matrix(design)
    N <- nrow(design)
    nonpar <- FALSE
  }
}

```

```

if (verbose) {
  cat("Finding maximum credible level at", N, "points using",
      M, "posterior draws.\n")
}

method <- method[1]
if (method == 'asymptotic') {
  m <- numeric(N)
  s <- numeric(N)
} else if (method == 'quantile') {
  Fxt <- Gxt <- numeric(N)
} else {
  warning("method must be one of 'asymptotic' or 'quantile'. Given: ",
          method)
  return(sim.cred.band)
}

if (z.store[1] == "ram") {
  z.store <- matrix(0, nrow=M, ncol=N)
  recompute.z <- FALSE
} else if (z.store[1] == "disk") {
  if (!require(ff)) {
    warning("Package 'ff' required to use function 'credsubs.level'",
            "with option z.store='disk'")
    return(sim.cred.band)
  }
  z.store <- ff(0, dim=c(M, N))
  recompute.z <- FALSE
} else if (z.store[1] == "recompute") {
  recompute.z <- TRUE
} else {
  warning("option z.store must be one of 'ram', 'recompute', or 'disk'.",
          "Given: ", z.store[1])
  return(rep(NA, N))
}

```

```

W <- rep(-Inf, M)
max.i <- numeric(M)
m <- s <- est <- numeric(N)
q <- sgn <- rep(NA, N)
if (!is.null(W.probs)) {
  W.quantile <- matrix(Inf, nrow=length(W.probs), ncol=N)
  rownames(W.quantile) <- paste0(W.probs * 100, "%")
}
test.set <- 1:N

for (i in 1:N) {

  if (verbose & i %% 100 == 0) {
    cat("prep", i, "/", N, "\n")
  }

  if (nonpar) {
    fx <- params[, i]
  } else {
    fx <- FUN(design[i, , drop=FALSE], params)
  }

  if (method == 'asymptotic') {
    m[i] <- mean(fx)
    s[i] <- sd(fx)
    est[i] <- m[i]
    z <- abs(fx - m[i]) / s[i]
  } else {
    est[i] <- median(fx)
    Fxt[i] <- mean(fx <= threshold)
    Gxt[i] <- mean(fx < threshold)
    Fx <- to.Fx(fx)
    Gx <- 1 - to.Fx(-fx)
    z <- pmax(1-Fx, Gx)
  }
}

```

```

if (!recompute.z) {
  z.store[, i] <- z
}

sgn <- ifelse(est > threshold, 1,
              ifelse(est < threshold, -1, 0))

max.i <- ifelse(z > W, i, max.i)
W <- pmax(W, z)
}

first.max.i <- max.i
prev.q <- 0
recompute.m <- numeric(0)

while (length(test.set) > 0) {
  if (verbose & (N - length(test.set)) %% 1 == 0) {
    cat("compute", N - length(test.set), "/", N,
        "recompute", length(recompute.m), "\n")
  }

  # Update W values
  if (length(recompute.m) > 0) {
    if (recompute.z) {
      if (method == 'asymptotic') {
        rcm <- recompute.m
      } else {
        rcm <- 1:M
      }
    }
    if (nonpar) {
      fx <- params[rcm, test.set, drop=FALSE]
    } else {
      fx <- t(FUN(X[test.set, , drop=FALSE],
                  params[rcm, , drop=FALSE]))
    }
  }
}

```

```

if (method == 'asymptotic') {
  Z <- t(abs(t(fx) - m[test.set]) / s[test.set])
} else { # method == quantile
  Fx <- t(apply(fx[recompute.m, , drop=FALSE], 1,
              function(x, m) {
                colMeans(t(t(m) <= x))
              }, m=fx))
  Gx <- 1-t(apply(-fx[recompute.m, , drop=FALSE], 1,
                function(x, m) {
                  colMeans(t(t(m) <= x))
                }, m=-fx))
  Z <- pmax(1-Fx, Gx)
}
} else {
  Z <- z.store[recompute.m, test.set, drop=FALSE]
}

max.i[recompute.m] <- apply(Z, 1, function(z) { test.set[which.max(z)] })
W[recompute.m] <- apply(Z, 1, max)
}

# Update empirical distribution of W
FW <- ecdf(W)
if (!is.null(W.probs)) {
  W.quantile[, test.set] <- quantile(W, probs=W.probs)
}

# Compute adjusted p-values
if (method == 'asymptotic') {
  lower <- 1 - FW((m[test.set] - threshold) / s[test.set])
  upper <- 1 - FW((threshold - m[test.set]) / s[test.set])
} else {
  lower <- 1-FW(1 - Fxt[test.set])
  upper <- 1-FW(Gxt[test.set])
}
}

```

```

p <- pmin(lower, upper)

if (step.down) {
  lowest.p.i <- test.set[which.min(p)]
  q[lowest.p.i] <- prev.q <- max(prev.q, p[which(test.set == lowest.p.i)])
} else {
  q <- p
  break
}

# these are the MCMC draws of W that change
# when the selected covariate point is removed

recompute.m <- which(max.i == lowest.p.i)

test.set <- test.set[which(test.set != lowest.p.i)]

i <- N - length(test.set)
if (verbose && i %% 100 == 0) {
  cat("step", i, "/", N, "(", length(recompute.m), "/", M, ")\n")
}
}

max.cred <- 1-q
attr(max.cred, "sign") <- sgn
if (!is.null(W.probs)) {
  attr(max.cred, "W.quantile") <- W.quantile
}
attr(max.cred, "threshold") <- threshold
attr(max.cred, "step.down") <- step.down
attr(max.cred, "first.max.i") <- first.max.i
class(max.cred) <- "credsubs.level"

max.cred
}

```

```
build.shiny.object <- function(credsubs.level, cov.space, location=".") {  
  
  # Validate input  
  stopifnot(length(credsubs.level) > 0)  
  stopifnot(is.numeric(credsubs.level))  
  stopifnot(!is.null(attr(credsubs.level, "sign")))  
  stopifnot(length(credsubs.level) == length(attr(credsubs.level, "sign")))  
  
  stopifnot(is.data.frame(cov.space))  
  stopifnot(length(credsubs.level) == nrow(cov.space))  
  
  # Save file  
  cat(paste0("Saving shiny object as ",  
            location, "/credsubs-shiny.RData", "\n"))  
  cat("Place file in same directory as shiny app.\n")  
  save(credsubs.level, cov.space,  
       file=paste0(location, "/credsubs-shiny.RData"))  
}
```

C.2 Shiny Calculator App

C.2.1 server.R

```

library(shiny)

shinyServer(function(input, output) {
  # should contain data frame 'cov.space' with rows representing
  # tested points in covariate space,
  # and vector 'credsubs.level' of maximum credible level at which
  # each covariate point is rejected, along with attribute 'sign'
  # which is a vector indicating whether the covariate point would be
  # in the exclusive group (1) or the complement of the inclusive group (-1)
  load("credsubs-shiny.RData")

  for (j in 1:length(cov.space)) {
    if (!is.factor(cov.space[, j]) &&
        (is.null(attr(cov.space[, j], "scaled:scale")) ||
         is.null(attr(cov.space[, j], "scaled:center")))) {
      attr(cov.space[, j], "scaled:scale") <- 1
      attr(cov.space[, j], "scaled:center") <- 0
    }
  }
}

output$predictors <- renderUI({
  inputs <- list()
  for (predictor in colnames(cov.space)) {
    if (is.factor(cov.space[, predictor])) {
      inputs[[predictor]] <-
        selectInput(predictor,
                    predictor,
                    unique(cov.space[, predictor])
                  )
    }
  }
})

```



```

if (length(matches) == 0) {
  return("Predictor combination untested.")
} else {
  return(paste0("A conclusion of <b>",
    ifelse(attr(credsubs.level, "sign")[matches] == 1,
      "<font color=\"#00AA00\">benefit</font>",
      "<font color=\"#AA0000\">no benefit</font>"),
    "</b> for patients with the above predictive covariate profile",
    "may be made at a maximum credible level of <b>",
    round(credsubs.level[matches] * 100, digits=2),
    "%</b>. ",
    "At higher credible levels, no conclusion may be drawn."))
  }
})
})

```

C.2.2 ui.R

```

shinyUI(fluidPage(

  titlePanel("Credible Subgroups Calculator"),

  uiOutput("predictors"),

  h3("Result"),
  htmlOutput("status")
))

```

Appendix D

Acronyms and Symbols

D.1 Acronyms

ATE	average treatment effect
PTE	personalized treatment effect
AD	Alzheimer's disease
SOC	standard of care
DRate	rate of decline
HPD	highest posterior density
RCS	restricted covariate space
PB	purely Bayesian
BART	Bayesian additive regression trees

D.2 Symbols

$a, \dots, z, \alpha, \dots, \omega$	scalars
$\mathbf{a}, \dots, \mathbf{z}, \boldsymbol{\alpha}, \dots, \boldsymbol{\omega}$	vectors
$\mathbf{A}, \dots, \mathbf{Z}$	matrices
A, \dots, Z	sets
$\mathcal{A}, \dots, \mathcal{Z}$	collections of sets

\mathbf{x}	prognostic covariate vector
\mathbf{z}	predictive covariate vector
t	treatment indicator
Y	outcome
μ_Z	probability law for Z
E	expected value
Δ	average treatment effect
$\Delta(\mathbf{z})$	personalized treatment effect at \mathbf{z}
B	benefiting subgroup
D	exclusive credible subgroup
S	inclusive credible subgroup
