



4-2001

A Comparison of Methods For Detection of Qualitative Interaction In Multicenter Trials

Boyd Jay Hanson
Western Michigan University

Follow this and additional works at: <https://scholarworks.wmich.edu/dissertations>



Part of the Design of Experiments and Sample Surveys Commons

Recommended Citation

Hanson, Boyd Jay, "A Comparison of Methods For Detection of Qualitative Interaction In Multicenter Trials" (2001). *Dissertations*. 1346.

<https://scholarworks.wmich.edu/dissertations/1346>

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact wmu-scholarworks@wmich.edu.



**A COMPARISON OF METHODS FOR DETECTION OF QUALITATIVE
INTERACTION IN MULTICENTER TRIALS**

by

Boyd Jay Hanson

**A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Department of Mathematics and Statistics**

**Western Michigan University
Kalamazoo, Michigan
April 2001**

A COMPARISON OF METHODS FOR DETECTION OF QUALITATIVE INTERACTION IN MULTICENTER TRIALS

Boyd Jay Hanson, Ph.D.

Western Michigan University, 2001

This research evaluated and compared three methods for the detection of qualitative treatment-by-center interaction proposed by Azzalini and Cox, Gail and Simon and Ciminera et al., through the analysis of simulated data for multicenter studies of two and three centers with two treatments. The effect of unequal sample size and the presence of an overall treatment effect on characteristics of the methods were examined.

The approach, underlying assumptions and theory of the three methods differ. For this study, they were adapted to establish a common basis for evaluation, thus allowing a meaningful comparison of the methods. For the test presented by Azzalini and Cox, this included deriving an approximate method and an exact method to allow for unequal sample sizes.

These methods were also compared to a common, ad-hoc method of identifying qualitative interaction, i.e. assessing the signs of the treatment effects by center. Two tests of overall interaction: the ANOVA test of interaction and the H statistic proposed by Gail and Simon were also examined.

Each method was further evaluated in a two-stage testing system, serving as a preliminary test to determine if the treatment-by-center interaction term should be included in the final analysis model.

The results indicate that the test for qualitative interaction proposed by Gail and Simon is the recommended method for detecting qualitative interaction. The error rates for patterns not exhibiting qualitative interaction are consistently lowest for this method. The second recommended choice would be the exact method of Azzalini and Cox.

This study did not provide a good evaluation of two-stage testing, except for the cases with equal sample sizes. For those cases, two-stage testing with one of the recommended methods was preferable to using a final analysis model without the interaction term.

This study was not designed to evaluate the effect that unequal sample size and inclusion of the treatment-by-center interaction term in the final model would have on the power of the test of overall treatment difference. However, the results show that in simulations with a high degree of imbalance, a study designed to have 80% power may have only 52%.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

UMI Number: 3007019

UMI[®]

UMI Microform 3007019

Copyright 2001 by Bell & Howell Information and Learning Company.

All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Copyright by
Boyd Jay Hanson
2001

ACKNOWLEDGMENTS

I thank family members and many friends and colleagues, whose support has sustained me throughout the lengthy process of obtaining this degree. I appreciate the financial and other material support from The Upjohn Company (now part of Pharmacia, Inc.) and Serono, Inc.

The Chair of my Committee, Mike Stoline, has provided valuable guidance in the formulation of the research and the completion of the dissertation. His patience and his adaptability to changing circumstances and ideas made the completion of this project a possibility. Dal Kratzer encouraged me to solidify my statistical knowledge, which led me to pursue this degree. He has been a valuable mentor and support throughout the process. Joe McKean provided valuable assistance, support and encouragement. Robert Buck, Dan Mihalko and Steve Francom provided reviews and discussion of the ideas and content of the dissertation that enriched the results.

Many colleagues at Upjohn and Serono willingly shared their time and ideas to assist me in the process.

My wife has patiently supported me and encouraged me through the completion of five degrees. I promise her that this will be the last one. I thank my children for their inspiration and support.

The completion of this doctoral program is a testament that prayers are heard and answered.

Boyd Jay Hanson

TABLE OF CONTENTS

ACKNOWLEDGMENTS	ii
LIST OF TABLES	xi
LIST OF FIGURES.....	xix
CHAPTER	
I. INTRODUCTION.....	1
II. ANALYSIS OF MULTICENTER TRIALS	4
2.1 The Multicenter Trial.....	4
2.2 Specification of the Analysis Model.....	5
2.3 Treatment-by-Center Interaction.....	5
2.4 Preliminary Test of the Treatment-by-Center Interaction.....	11
2.5 Type II or Type III Analysis.....	15
2.6 Analytical Studies Addressing the Analysis of Multicenter Trials	20
2.7 Synthesis	23
III. CURRENT METHODS FOR IDENTIFYING QUALITATIVE INTERACTION.....	25
3.1 Introduction.....	25
3.2 Definition of Terminology.....	26
3.2.1 Quantitative Interaction.....	27
3.2.2 Qualitative Interaction.....	27
3.3 Estimation of Parameters.....	29
3.4 Specification of the Analysis of Variance Model	31

Table of Contents—continued

CHAPTER

3.5	Ciminera, Heyse, Nguyen and Tukey	34
3.6	Gail and Simon	40
3.7	Azzalini and Cox	44
3.8	Evaluation and Comparison of Methods -- Previous Work.....	49
3.9	Summary.....	49
IV.	EVALUATION OF CURRENT METHODS OF IDENTIFYING OF QUALITATIVE INTERACTION	50
4.1	Detection of Interaction	50
4.2	Comparison of the Underlying Assumptions of the Methods ...	50
4.2.1	Ciminera, Heyse, Nguyen and Tukey	50
4.2.2	Gail and Simon	51
4.2.3	Azzalini and Cox	52
4.3	Establishing a Common Basis for Comparison of the Three Methods	53
4.4	Extension of the Azzalini and Cox Method to the Case of Unequal Sample Sizes Across Centers	55
4.4.1	Exact Approach.....	56
4.4.2	Approximate Approach.....	59
4.5	Adaptation of the Methods of Ciminera et al. and Gail and Simon to Incorporate the Variance Estimate From Analysis of Variance	62
4.6	Examination of the Signs of the Treatment Effects in the Raw Data	65
4.7	Example.....	65
4.7.1	Introduction and Results of the Analysis of Variance....	66

Table of Contents—continued

CHAPTER

4.7.2	Signs of the Treatment Effects in the Raw Data.....	68
4.7.3	Azzalini and Cox Exact Method.....	68
4.7.4	Azzalini and Cox Approximate Method.....	69
4.7.5	Gail and Simon	70
4.7.6	Ciminera et al.....	73
4.7.7	Summary.....	75
V.	DESCRIPTION OF SIMULATION PROCEDURES FOR COMPARING METHODS OF IDENTIFYING QUALITATIVE INTERACTION.....	78
5.1	Introduction.....	78
5.2	Specification of Simulated Datasets	79
5.3	Evaluation of Simulated Datasets	81
5.3.1	Simulation Procedures	81
5.3.2	Tests of Overall Interaction.....	83
5.3.3	Methods for the Detection of Qualitative Interaction	84
5.3.4	Test of Non-inferiority	85
5.3.5	Two-stage Test Results	86
5.4	The General Linear Mixed Model	87
5.5	Simulation of Trials With Two Centers.....	89
5.5.1	Patterns With a Difference Between Treatments (Overall Treatment Effect).....	89
5.5.2	Patterns With No Difference Between Treatments (No Overall Treatment Effect).....	91
5.5.3	Difference Parameters.....	92

Table of Contents—continued

CHAPTER		
	5.5.4 Sample Sizes	96
5.6	Simulation of Trials With Three Centers.....	97
	5.6.1 Patterns With a Difference Between Treatments (Overall Treatment Effect).....	98
	5.6.2 Patterns With No Difference Between Treatments (No Overall Treatment Effect).....	100
	5.6.3 Sample Sizes	101
VI.	RESULTS FROM SIMULATED DATA FOR TWO CENTERS	107
	6.1 Introduction.....	107
	6.2 Case 2.1: 32 Patients at Center 1 and 32 Patients at Center 2 ..	108
	6.2.1 General Results	108
	6.2.2 Tests of Overall Interaction.....	108
	6.2.3 Methods for the Detection of Qualitative Interaction	109
	6.2.4 Test of Non-inferiority	114
	6.2.5 Two-stage Test Results	116
	6.2.6 Evaluation of Redundant Patterns.....	120
	6.3 Case 2.2: 43 Patients at Center 1 and 21 Patients at Center 2 ..	120
	6.3.1 Tests of Overall Interaction.....	121
	6.3.2 Methods for the Detection of Qualitative Interaction	122
	6.3.3 Test of Non-inferiority	124
	6.3.4 Two-stage Test Results	126
	6.4 Case 2.3: 21 Patients at Center 1 and 43 Patients at Center 2 ..	129
	6.4.1 Tests of Overall Interaction.....	129

Table of Contents—continued

CHAPTER

6.4.2	Methods for the Detection of Qualitative Interaction	130
6.4.3	Test of Non-inferiority	132
6.4.4	Two-stage Test Results	134
6.5	Comparison of Case 2.2 and Case 2.3	136
6.6	Case 2.4: 54 Patients at Center 1 and 10 Patients at Center 2 ..	140
6.6.1	Tests of Overall Interaction.....	141
6.6.2	Methods for the Detection of Qualitative Interaction	142
6.6.3	Test of Non-inferiority	144
6.6.4	Two-stage Test Results	149
6.7	Case 2.5: 10 Patients at Center 1 and 54 Patients at Center 2 ..	149
6.7.1	Tests of Overall Interaction.....	149
6.7.2	Methods for the Detection of Qualitative Interaction	151
6.7.3	Test of Non-inferiority	153
6.7.4	Two-stage Test Results	154
6.8	Comparison of Case 2.4 and Case 2.5	157
6.9	Discussion.....	161
6.9.1	Test of Overall Interaction and Methods for the Detection of Qualitative Interaction	161
6.9.2	Test of Non-inferiority	164
6.9.3	Two-stage Test Results	166
VII.	RESULTS FROM SIMULATED DATA FOR THREE CENTERS	169
7.1	Introduction.....	169

Table of Contents—continued

CHAPTER

7.2	Case 3.1: 22 Patients at Center 1, 21 Patients at Center 2 and 21 Patients at Center 3	170
7.2.1	General Results	170
7.2.2	Tests of Overall Interaction.....	171
7.2.3	Methods for the Detection of Qualitative Interaction	173
7.2.4	Test of Non-inferiority	176
7.2.5	Two-stage Test Results	178
7.3	Case 3.2: 39 Patients at Center 1, 13 Patients at Center 2 and 12 Patients at Center 3	182
7.3.1	Tests of Overall Interaction.....	183
7.3.2	Methods for the Detection of Qualitative Interaction	184
7.3.3	Test of Non-inferiority	187
7.3.4	Two-stage Test Results	188
7.4	Case 3.3: 13 Patients at Center 1, 39 Patients at Center 2 and 12 Patients at Center 3	192
7.4.1	Tests of Overall Interaction.....	192
7.4.2	Methods for the Detection of Qualitative Interaction	194
7.4.3	Test of Non-inferiority	196
7.4.4	Two-stage Test Results	196
7.5	Case 3.4: 13 Patients at Center 1, 12 Patients at Center 2 and 39 Patients at Center 3	200
7.5.1	Tests of Overall Interaction.....	201
7.5.2	Methods for the Detection of Qualitative Interaction	201
7.5.3	Test of Non-inferiority	204

Table of Contents—continued

CHAPTER

7.5.4	Two-stage Test Results	206
7.6	Case 3.5: 29 Patients at Center 1, 29 Patients at Center 2 and 6 Patients at Center 3	209
7.6.1	Tests of Overall Interaction.....	209
7.6.2	Methods for the Detection of Qualitative Interaction....	211
7.6.3	Test of Non-inferiority	213
7.6.4	Two-stage Test Results	214
7.7	Case 3.6: 29 Patients at Center 1, 6 Patients at Center 2 and 29 Patients at Center 3	218
7.7.1	Tests of Overall Interaction.....	218
7.7.2	Methods for the Detection of Qualitative Interaction....	219
7.7.3	Test of Non-inferiority	221
7.7.4	Two-stage Test Results	223
7.8	Case 3.7: 6 Patients at Center 1, 29 Patients at Center 2 and 29 Patients at Center 3	226
7.8.1	Tests of Overall Interaction.....	226
7.8.2	Methods for the Detection of Qualitative Interaction....	228
7.8.3	Test of Non-inferiority	230
7.8.4	Two-stage Test Results	231
7.9	Discussion.....	235
7.9.1	Tests of Overall Interaction and Methods for the Detection of Qualitative Interaction	235
7.9.2	Test of Non-inferiority	240
7.9.3	Two-stage Test Results	243

Table of Contents—continued

CHAPTER

VIII. CONCLUSIONS AND FUTURE WORK.....	246
8.1 Conclusions.....	246
8.2 Future Work.....	249

APPENDIX

Simulation Programs.....	251
BIBLIOGRAPHY	319

LIST OF TABLES

1. Analysis of Variance Table for a Multicenter Trial	32
2. Comparison of Some Important Assumptions of the Three Methods for Identifying Qualitative Interaction	55
3. Summary Statistics for Patient Response.....	67
4. Tests of Fixed Effects	67
5. Results of the Ciminera et al. Procedure.....	77
6. Description of Two-Center Simulation Patterns, Interaction Type, Predicted Mean Responses and Effect Sizes for Each Treatment-by- Center Cell	90
7. Description of Two-Center Simulation Patterns, Interaction Type, Predicted Mean Responses and Effect Sizes for Each Treatment-by- Center Cell	92
8. Sample Sizes for Each Treatment Group for Simulations of Interaction Patterns for Two Centers	98
9. Description of Three-Center Simulation Patterns, Interaction Type and Predicted Mean Responses for Each Treatment-by-Center Cell.....	100
10. Description of Three-Center Simulation Patterns, Interaction Type and Predicted Mean Responses for Each Treatment-by-Center Cell.....	101
11. Sample Sizes for Each Treatment Group for Simulations of Interaction Patterns for Three Centers.....	106
12. Means and Variances by Treatment and Center of Two-Center Simulated Data.....	108
13. Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 2.1	110
14. Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 2.1	112

List of Tables—continued

15.	Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers for Case 2.1	115
16.	Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 2.1	117
17.	Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 2.1	118
18.	Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 2.2.....	121
19.	Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 2.2	123
20.	Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers for Case 2.2	125
21.	Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 2.2	127
22.	Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 2.2.....	128
23.	Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 2.3.....	130
24.	Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 2.3	131
25.	Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers for Case 2.3	133
26.	Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 2.3	134

List of Tables—continued

27.	Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 2.3.....	136
28.	Differences (Case 2.2 – Case 2.3) in Treatment Means for Type III and Type II Models.....	138
29.	Differences (Case 2.2 – Case 2.3) in Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction.	139
30.	Difference (Case 2.2 – Case 2.3) in Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction	140
31.	Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 2.4.....	141
32.	Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 2.4.....	143
33.	Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers for Case 2.4.....	145
34.	Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 2.4	147
35.	Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 2.4.....	148
36.	Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 2.5.....	150
37.	Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 2.5	151
38.	Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers for Case 2.5	153

List of Tables—continued

39.	Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 2.5	155
40.	Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 2.4.....	156
41.	Differences (Case 2.4 – Case 2.5) in Treatment Means for Type III and Type II Models.....	158
42.	Differences (Case 2.4 – Case 2.5) in Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction.	159
43.	Differences (Case 2.4 – Case 2.5) in Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction	160
44.	Minimum and Maximum Percentages of Simulations With Significant Tests or Substantial Evidence of Overall Interaction or Qualitative Interaction for Two-Center Simulations	162
45.	Minimum and Maximum Percentages of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers	165
46.	Minimum and Maximum Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction.....	167
47.	Means and Variances by Treatment and Center of Three-Center Simulated Data.....	170
48.	Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.1.....	172
49.	Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.1	174
50.	Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.1	178

List of Tables—continued

51.	Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 3.1	180
52.	Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 3.1	181
53.	Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.2.....	183
54.	Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.2	185
55.	Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.2	188
56.	Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 3.2	189
57.	Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 3.2.....	191
58.	Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.3.....	193
59.	Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.3	194
60.	Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.3	197
61.	Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 3.3	198
62.	Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 3.3.....	200

List of Tables—continued

63.	Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.4.....	202
64.	Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.4	203
65.	Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.4	205
66.	Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 3.4	206
67.	Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 3.4.....	208
68.	Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.5.....	210
69.	Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.5	211
70.	Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.5	213
71.	Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 3.5	215
72.	Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 3.5.....	217
73.	Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.6.....	219
74.	Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.6	220

List of Tables—continued

75. Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.6	222
76. Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 3.6	224
77. Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 3.6.....	225
78. Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.7.....	227
79. Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.7	229
80. Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.7	230
81. Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 3.7	232
82. Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 3.7.....	234
83. Minimum and Maximum Percentages of Simulations With Significant Tests or Substantial Evidence of Overall Interaction or Qualitative Interaction for Three-Center Simulations	236
84. Sample Sizes by Center for “Strong” and “Weak” Cases of Qualitative Interaction. Effect Sizes for Each Pattern at Each Center are Also Indicated.....	240
85. Differences in Qualitative Interactions Detected in Cases of “Strong” Qualitative Interaction and Cases of “Weak” Qualitative Interaction for Three-Center Simulated Data.....	241

List of Tables—continued

86.	Minimum and Maximum Percentages of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers	242
87.	Minimum and Maximum Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction.....	244

LIST OF FIGURES

1. Mean Responses for Two Treatments (Treatment 1 —, Treatment 2 --) at Three Centers with No Interaction.	28
2. Mean Responses for Two Treatments (Treatment 1 —, Treatment 2 --) at Three Centers with Quantitative Interaction.	29
3. Mean Responses for Two Treatments (Treatment 1 —, Treatment 2 --) at Three Centers with Qualitative Interaction.	30
4. Simulated Interaction Patterns for Two Centers: Patterns 1 - 3 (Treatment 1 —, Treatment 2 --).	93
5. Simulated Interaction Patterns for Two Centers: Patterns 4 - 6 (Treatment 1 —, Treatment 2 --).	94
6. Simulated Interaction Patterns for Two Centers: Patterns 6 - 9 (Treatment 1 —, Treatment 2 --).	95
7. Simulated Interaction Patterns for Two Centers: Patterns 10 - 11 (Treatment 1 —, Treatment 2 --).	96
8. Difference Parameters for Interaction Patterns for Two Centers	97
9. Simulated Interaction Patterns for Three Centers: Patterns 1 - 3 (Treatment 1 —, Treatment 2 --).	102
10. Simulated Interaction Patterns for Three Centers: Patterns 4 - 6 (Treatment 1 —, Treatment 2 --).	103
11. Simulated Interaction Patterns for Three Centers: Patterns 7 - 9 (Treatment 1 —, Treatment 2 --).	104
12. Simulated Interaction Patterns for Three Centers: Patterns 10 - 11 (Treatment 1 —, Treatment 2 --).	105

CHAPTER I

INTRODUCTION

The conduct of clinical trials to establish the efficacy and safety of investigational compounds is an important aspect of modern pharmaceutical research. These trials are generally carried out at several locations, or study centers, with each center using a common protocol. This experimental design is called a multicenter trial. Although this design is generally used, the analysis of data from multicenter trials is the subject of controversy among clinicians and statisticians. The multicenter trial and some of the issues related to the analysis of data from these trials is presented in Chapter II.

One area of dispute is the detection and management of treatment-by-center interaction. This topic is further examined in Chapter III. True treatment-by-center interaction arises when the effect of the treatment on similarly treated patients differs among centers. However, among-center differences in the mean treatment effect may be present in multicenter trials due to statistical variability among centers or due to the presence of true treatment-by-center interaction. Treatment-by-center interaction can be classified as quantitative interaction or qualitative interaction. Interaction is defined as quantitative when treatment differences among centers are all in the same direction (i.e. positive or negative), but differ in degree. For example, the mean total blood cholesterol level of patients treated with Treatment 1 may be lower than mean levels of patients treated with Treatment 2 at all centers. However, the difference between the two mean levels may differ numerically among the centers. Qualitative interactions arise when the treatment differences among centers differ in direction.

For example, the mean total blood cholesterol level of patients treated with Treatment 1 may be lower than mean levels of patients treated with Treatment 2 at most centers. But at one or more centers, the mean level of patients treated with Treatment 1 may be higher than mean levels of patients treated with Treatment 2. The detection of qualitative interaction may be especially important to the interpretation of results from a trial. Three statistical methods have been proposed that could be used to detect the presence of qualitative interaction. The presentation of these methods is the focus of Chapter III.

The operating characteristics of the three methods of detecting qualitative interaction have not been determined or compared. The utility of the methods could be enhanced if their characteristics were more completely understood. The objective of this research is to evaluate and compare these methods under defined conditions. In Chapter IV, the underlying assumptions of the three methods are compared and a common basis for comparison of the methods is established. An example of the analysis of a simple multicenter trial with the calculation of the results using each of the methods is also presented.

The methodology chosen for the evaluation and comparison of the methods of identifying qualitative interaction was the analysis of simulated datasets prepared with characteristics that allowed the evaluation and comparison of the methods under specified conditions. The simulation procedures used to compare the three methods are described in Chapter V.

We began our research with the simplest form of qualitative interaction; we simulated trials with two treatments and two centers. We also evaluated designs with two treatments and three centers.

The results of the simulations for the two-center designs are presented in Chapter VI and the results of the three-center designs in Chapter VII. Overall conclusions are presented in Chapter VIII, as well as some suggestions for future research.

CHAPTER II

ANALYSIS OF MULTICENTER TRIALS

2.1 The Multicenter Trial

The multicenter trial is an important design for pharmaceutical research studies. A multicenter trial is a study that is conducted at two or more locations, using a common protocol at each location. The ICH (International Conference on Harmonization) Guidance on Statistical Principles for Clinical Trials (International Conference on Harmonisation, 1998, p. 49589) states:

Multicenter trials are carried out for two main reasons. First, a multicenter trial is an accepted way of evaluating a new medication more efficiently. Under some circumstances, it may present the only practical means of accruing sufficient subjects to satisfy the trial objective within a reasonable timeframe...

Second, a trial may be designed as a multicenter (and multi-investigator) trial primarily to provide a better basis for the subsequent generalization of its findings.

These advantages of multi-center trials are very important and have led to the widespread acceptance of the design. However, the use of this experimental design presents a number of statistical issues.

As Fleiss (1986, p. 267) stated in his 1986 review article,

Given the relatively long history of multiclinic trials, and given the relatively large literature on the planning and execution of such studies, there is a striking paucity of articles and chapters in books on the analysis of the resulting data. This dearth in the literature is especially striking because uncertainty and controversy exist concerning most aspects of analysis.

Although some additional authors have treated the subject in the last decade, the subject is still embroiled in controversy and many questions are still unanswered.

2.2 Specification of the Analysis Model

One of the most important issues in the analysis of data from multi-center trials is the specification of the model to be used to analyze the data. Bancroft (1964, p. 427) indicates that Fisher was of the opinion that “an appropriate statistical-mathematical model, to be used in describing an observation in an investigation, should be determined in advance by an investigator.” Lewis (1995, p. 132) states, “The pre-specification of the anticipated analysis in the protocol of a clinical trial is now an accepted standard in the pharmaceutical industry, and this is well reflected in regulatory guidelines. There only remains debate about the detail.” He further discusses the benefits of model pre-specification, as well as the apprehensions felt by some statisticians. One of the benefits he discusses is the avoidance of bias arising from the selection of the most favorable analysis.

However, the specification of the model in a multicenter trial is problematic because of several issues.

2.3 Treatment-by-Center Interaction

An initial issue in the construction of the analysis model is the inclusion of the treatment-by-center term in the primary analysis model. Two alternative analysis models have been proposed for the analysis of data from a multicenter trial. These models will be described in more detail in Chapter III; however, each model will be introduced here.

The response data from a multicenter trial can be described with a model that includes a term for the treatment-by-center interaction. This model is often called a Type III model and has the following form:

$$(2.1) \quad y_{ijk} = \mu + B_i + C_j + BC_{ij} + e_{ijk}, \text{ where}$$

$i = 1, \dots, b$ (number of treatments),

$j = 1, \dots, c$ (number of centers),

$k = 1, \dots, n_{ij}$ (number of patients assigned to treatment i at center j),

y_{ijk} is the measured response of patient k assigned to treatment i at center j ,

μ is the overall mean response,

B_i is the fixed effect of the i th treatment (the treatment mean of treatment i is $\mu_i = \mu + B_i$),

C_j is the fixed effect of the j th center,

BC_{ij} is the fixed interaction effect of the i th treatment and the j th center and

e_{ijk} is the random error associated with patient k assigned to treatment i at center j .

Alternatively, the response data from a multicenter trial can be described with a model that does not include a term for the treatment-by-center interaction. This model is often called a Type II model and has the following form:

$$(2.2) \quad y_{ijk} = \mu + B_i + C_j + e_{ijk},$$

with terms defined as above.

Fleiss (1986, p. 272) wrote, “The most challenging questions in the analysis of data from a multi-center trial are how to carry out the analysis when there is treatment-by center interaction, and prior to that, how to ascertain whether such interaction exists.” An additional level of uncertainty may be added to these questions by regulatory guidance documents, which may make recommendations for testing the significance of the interaction before combining results across centers.

However they “provide neither warning nor reassurance as to the significance level at which the test should be performed, nor do they help determine the degree of inconsistency of results across clinics that would make the combination of results unsuitable.”(Fleiss, 1986, p. 272)

Gallo (1998, p. 2) discusses three common industry approaches to the inclusion of the treatment-by-center interaction in the primary analysis model. They are:

- (a) include this interaction term in the model;
- (b) do not include it (possibly addressing the possibility of meaningful interaction in supplemental analyses);
- (c) a hybrid of approaches (a) and (b), in which the interaction is initially included, but a preliminary significance test for the interaction is performed; if the term is not significant (a level of 0.10 is often used), it is removed prior to performing the test for treatment.

He indicates that, “In the U.S. pharmaceutical industry, a common practice (quite possibly the majority practice) is (a)...”

Gould (1998, p. 1779) concurs, “...the conventional analysis, at least in the U.S.A., is a fixed effects ANOVA with terms for centres, treatments, and centre x treatment interactions. Whether this represents the best way to approach the problem remains to be seen.”

In his discussion of the determination of the primary analysis model, Gallo (1998, p. 2) proposes that the treatment-by-center interaction should be treated as an exploratory, rather than an explanatory factor. He states that “we rarely undertake a trial with a clear expectation regarding the nature of the different effects expected in different centers.” Källén (1997, p. 932) suggests that,

The interaction test should not be forgotten, but considered an independent test which is done primarily to assess possible differences in study conduct between centers. This is one of a number of different

tests that should be done in order to detect outliers, subgroups with special response profiles, and the like.

The ICH guideline “Structure and Contents of Clinical Study Reports”

(International Conference on Harmonisation, 1996, p. 37327) states,

Individual center results should be presented, however, where appropriate, e.g., when the centers have sufficient numbers of patients to make such an analysis potentially valuable, the possibility of qualitative or quantitative treatment-by-center interaction should be explored. Any extreme or opposite results among centers should be noted and discussed, considering such possibilities as differences in study conduct, patient characteristics, or clinical settings. Treatment comparison should include analyses that allow for centre differences with respect to response.

The interpretability of the treatment-by center interaction is an important consideration in the decision of whether or not to include the interaction term in the model. Several authors believe that some amount of treatment-by-center interaction is inevitable, especially quantitative interaction. Interaction is defined as quantitative when treatment differences among centers are all in the same direction (i.e. positive or negative), but differ in degree. Qualitative interactions arise when the treatment differences among centers differ in direction. These types of interactions will be further described and discussed in Chapter III.

Gail and Simon (1985) state that quantitative interactions are to be expected and may be artifacts of the scale of measurement. Snappin (1998, p. 433) illustrates that the presence or absence of quantitative interaction may depend on scale. He also notes that,

The very differences in patient demographics, medical conditions, and treatment milieu (e.g., certain concomitant medications or medical procedures might be more common in one region of the country than in another) that make the results of multiclinic trials more generalizable, however, also introduce the potential for the results to differ among clinics. When these differences affect the randomized

treatment regimens unequally, treatment-by-clinic interaction is introduced.

In addition, Snappin indicates that departures from the protocol and different criteria for evaluating the response at certain centers may result in treatment-by-clinic interaction. Jones et al. (Jones, B., Teather, D., Wang, J., & Lewis, J. A., 1998, p. 1769) also highlight the inevitability of some degree of treatment-by-center interaction. They state that, “it is likely that the treatment difference will not be the same in every centre. That is, the existence of some degree of treatment-by-centre interaction is almost certain.”

Källén (1997, p. 935) attributes some interaction to small centers because of the difficulty of reliably estimating the means. He also states that the

interaction is a consequence of biological variation in the population studied and/or differences in the conduct between medical centers, both of which are well known. If these were actual problems, clinical studies could not be done at all. Therefore a significant interaction should not come as a surprise, and must not invalidate the treatment comparison.

Senn (1998, pp. 1760-1761) points out that another source of apparent interaction is the low precision of center-specific treatment estimates. This leads to considerable chance variation between centers. He gives an example,

Consider a placebo controlled multi-centre trial with 80 per cent power in total at 5 per cent level (two-sided). Suppose that the true treatment effect is identically equal to the clinically relevant difference. It then follows that provided we have at least six centres, there is an odds on chance that at least one of them will show an ‘effect reversal’ (the placebo will appear superior).

This chance variation in the direction of the treatment effect can lead to confusion in the interpretation of the results.

Since some amount of observed treatment-by-center interaction seems inevitable, when does it become a problem? Gould (1998, p. 1780) states,

The definition of “interaction” is a central issue. If centres are viewed as fixed effects, then interaction can be defined as variation in the true treatment fixed effects over levels of the true centre fixed effects. If the *observed* variation in the treatment effects across centres exceeds what can be explained by sampling error alone, then either the model expressing what chance alone would predict is wrong, or some of the centres possess attributes affecting the actions of the treatments (true interaction) and a serious attempt must be made to identify these attributes. Which of these situations applies in any instance is a judgement call.

Jones et al. (1998, p. 1769) note that some argue that “the treatment-by centre interaction is usually of no interest and can be ignored. However, whether it is to be studied or ignored, it will inevitably have an influence on the estimated treatment effect, and this influence needs to be understood.”

Källén (1997, p. 929) feels that the interaction should not be an issue, based on clinical considerations, not statistical. He argues that,

The basic difference between the statistician and the clinician is that the latter studies patients, the former centers. But the clinician has a point—patients are being treated, not centers. This does not mean that the interaction should be ignored. It tells researchers that the population is heterogeneous and could make it possible to identify some important heterogeneity factors. But it does not invalidate conclusions based on average treatment effects!

Gould (1998, p. 1794) (who advocates a Bayesian approach to the analysis rather than the fixed frequentist approach discussed here) states that,

..it is not clear that this interaction has much meaning in the usual context of multi-centre trials as opposed to the factorial design circumstances for which it is intended. Regardless of the model, centres whose outcomes differ markedly from those of the bulk of the centres, that is, apparent outliers, need to be followed up to determine why, since the effects of treatments applied in a consistent manner to similar kinds of patients should be manifested consistently regardless of the centre in which the observations are made.

Lewis (1995, p. 132) writes that,

substantial treatment-by-centre interactions should be just as unlikely as any other sort of treatment interaction and are not our main interest. (In fact substantial treatment-by-centre interactions, if and when they exist, must be caused by something else, such as differences in specific patient characteristics or clinical conditions. Their interpretative value is purely as a possible indicator of something more valuable for future use.)

He continues (p. 132),

Even when a substantial and statistically significant treatment-by-centre interaction is detected, it should not be the end of the road. An interaction is a sign of a treatment difference, albeit a non-constant one and if the interaction is quantitative rather than qualitative, the problem may be containable. The effect is established; it only remains to put bounds on its size.

Källén (1997, p. 928) reminds his readers that if the interaction is tested for significance at a level of 10% that, “even if treatment differences never differ between centers, on the average, 10% of all studies will pose a problem, since the statistical test will turn out to be significant.”

2.4 Preliminary Test of the Treatment-by-Center Interaction

A common practice in the U.S. pharmaceutical industry is to include the interaction term in the model and conduct a two-stage testing procedure. At the first stage, the interaction term is tested for significance, usually at a level of 10%. The second stage is the test of significance of treatment effect.

The power of the test of interaction (first stage) is an issue. Gallo (1998, p. 2) notes that study designs rarely consider the sample size needed to provide adequate power to test the significance of the interaction. He further states that “we rarely undertake a trial with a clear expectation regarding the nature of different effects expected in different centers.” Lewis (1995, p. 131) reminds his readers,

It is vital to remember that the multi-centre design is nearly always adopted to accrue sufficient patients within a realistic time. The power to look at interactions is therefore bound to be small... Indeed, the decision to carry out a multi-centre trial requires the *assumption* that the treatment effect is similar from centre to centre, because we will be unable to test this adequately at the end of the study.

The ICH guidance on “Statistical Principles for Clinical Trials” (1998, p. 49589) also emphasizes the lack of power of the test, and recommends excluding the treatment-by center interaction from the model,

The statistical model to be adopted for the estimation and testing of treatment effects should be described in the protocol. The main treatment effect may be investigated first using a model that allows for center differences, but does not include a term for treatment-by-center interaction. If the treatment effect is homogeneous across centers, the routine inclusion of interaction terms in the model reduces the efficiency of the test for the main effects. In the presence of true heterogeneity of treatment effects, the interpretation of the main treatment effect is controversial.

Marked heterogeneity may be identified by graphical display of the results of individual centers or by analytical methods, such as a significance test of the treatment-by-center interaction. When using such a statistical significance test, it is important to recognize that this generally has low power in a trial designed to detect the main effect of treatment.

The use of a preliminary test for interaction is challenged by Jones et al. (1998), based on the results of a simulation study that examined the merits of conducting a preliminary test for interaction at a 10% significance level. They concluded that (p. 1776),

Pre-testing for treatment-by-centre interaction and then using a different estimator dependent on test outcome suffers from similar problems to pre-testing for a carry-over difference in cross-over trials. The performance of the pre-testing approach in this simulation study provides little evidence to recommend it.

However, Fleiss (1986, pp. 272-273) states,

Even though the chances are one in ten that there will be undue concern about interaction and perhaps even an inefficient analysis when interaction is not really there, the relatively high power that such a criterion provides when in fact there is interaction is reassuring to those who believe that clinical or demographic differences between the clinics' patients make interaction possible, and that loose controls and little or no monitoring of procedures at the individual clinics make interaction inevitable.

Fleiss recommends a test of interaction at a level of 10%.

Fleiss and other advocates of the two-stage procedure feel that if the preliminary test of the treatment-by center interaction does not support the existence of the interaction, then the interaction can safely be omitted from the analysis model. Without the interaction in the model, then the analysis model has simply the terms for treatment and center.

However, if the preliminary test of the treatment-by center interaction supports the existence of the interaction, the interpretation of the analysis in the presence of the interaction is controversial.

Fleiss (1986) describes two methods of analysis in the presence of heterogeneity. The first is the use of the Type III analysis, which he attributes to Yates (1934). The second method is one suggested by Overall (1979). However, Fleiss warns that this latter method discards data in order to obtain a consistent, statistically significant difference, which may damage the credibility of the final result. He concludes with an appeal (p. 274),

Yet other procedures have doubtless been proposed or applied when interaction is present, but the relevant literature, in which they are presented, analyzed, and criticized does not seem to exist. Perhaps this review article will provide an impetus for publication, debate, and, ultimately, consensus on how best to analyze the data from a multiclinic trial when treatment-by-clinic interaction exists.

Lin's (1999, p. 370) approach to overcoming the interpretation difficulties associated with the treatment-by-center interaction is to examine the population represented by the study. In the presence of a treatment-by-center interaction, the patient populations at each center cannot be assumed to be the same and there is no well-defined population as a basis for the analysis. However, he defines the "Study Represented Population (SRP)" which is a patient population that is "specific to the centers in the study, and to the proportions of patients recruited from each of the centers." However he notes that "one believes (or hopes) that conclusions drawn from SRP can be generalized to the target population specified in the protocol." He explains (p. 371) that the bias of the Type II and Type III means in estimating the true population mean can not be accurately determined.

If, at one extreme, all the centers are equally motivated, and use the same practice in recruiting study patients, then the proportions of patients enrolled at study centers become natural estimates of the proportions of the patient population the centers serve...At the other extreme, if all centers serve the same size of patients populations but have different motivations in recruiting patients, the result is that sample sizes at different centers differ substantially...In reality, different centers not only serve different sizes of patient population, but they are also not equally motivated. Usually we do not know exactly the extent of these differences and therefore have no assurance as to which of the two parameters is the better approximation.

He believes that the Type II analysis is valid, even in the presence of interaction (p. 372).

Because in *SRP*, the proportions of patients belonging to different centers are unknown constants and the weights estimate these proportions, these weights have to be sample related. While *SRP* usually is not exactly equal to the entire patient population targeted in the study protocol in a statistical sense, it is the patient population under the specific clinical trial settings and is the representation of the targeted patient population by the study. Therefore the weighted analysis is not only statistically valid and interpretable, it represents what has been provided from the clinical study.

Ciminera et al. (Ciminera, J. L., Heyse, J. F., Nguyen, H. H., & Tukey, J. W., 1993, p. 1043) advocate an alternative method only if qualitative interaction is present.

If there is substantial evidence of qualitative interaction, then a non-typical situation has emerged. The reason must be sought through discussions with the medical team and by exploratory analysis of the data. The combination of results in the presence of qualitative interaction should involve the treatment-by-centre-interaction mean square as the error term for the treatment effect. However this should not be done unless the evidence for qualitative interaction is “substantial”.

2.5 Type II or Type III Analysis

The choice of the appropriate statistic to use as the estimator of the treatment effect is another issue in the construction of the analysis model.

Generally, the parameter(s) of interest in a multicenter trial is the comparison across all centers of the effect of the study drug versus a control. At issue is the statistic to use as an estimate of the means and the differences of the treatment groups to be compared. Two popular parametric estimates are the unweighted or Type III mean and the weighted or Type II mean. The Type III mean is the “mean of means”; it is the mean of the treatment means from each location. It is termed unweighted because the center treatment means are not weighted by the number of patients at each center. On the other hand, the Type II mean uses the means from each center weighted by a function of the sample size of the treatment groups at each center. (Although functions of the variance are used as a weights in some analyses, the weights discussed in this study always refer to functions of the sample size.)

Senn (1998) compares the two estimators using the analogy of the US House of Representatives (Type II) and the US Senate (Type III). The Type II estimator

provides additional “votes” to centers with higher enrollment, whereas the Type III estimator gives equal “votes” to each center regardless of the sample size. In accordance with this analogy, the Type II method is sometimes labeled as the “one patient, one vote” method.

When sample sizes are equal across all centers, then the weighted and unweighted means are equal and the differences between the means are equal. However, it is often necessary to designate either the Type II or Type III as the primary analysis model. Arguments for and against the two methods have been made on the basis of both the appropriateness and the variance of the estimators.

As Gallo (1998) indicates, the general procedure in the United States has been to employ a Type III analysis. This is consistent with the 1988 recommendation of the US Food and Drug Administration (Center for Drug Evaluation and Research, Food and Drug Administration, U. S. Department of Health and Human Services, 1988) for New Drug Applications. Fleiss (1986) states that this is the analysis that makes sense when the treatment effects are not consistent across centers. Goldberg and Koury (1990) also warn that Type II estimators are not unbiased when there is treatment-by-center interaction.

Gallo (1998, p. 3) argues that “even if there is a treatment-by-center interaction, but there is nothing systematic about the relationship between center size and within center effect, then both approaches are based on essentially the same parameter.” Hence, the weighted and unweighted effects are equal. He illustrates with an example where there is an interaction, as well as unequal sample sizes, but the treatment effect is not correlated with center size. He concludes (p. 4) that when,

the weighted and unweighted effects are meaningfully different, it seems difficult to conclude that either is inherently ‘better’ (except from an efficiency standpoint, for which the answer is clear); the difference may be saying something important about the nature of the

interaction (e.g., a larger effect in large centers) which hopefully should be determined from a thorough examination of the data.

Since the estimators may be comparable, Gallo recommends using the more efficient, Type II, estimator. Efficiency of the Type II and Type III estimators will be discussed later.

Senn (1998, p. 1756) summarizes some of the arguments of proponents for each method.

The arguments made in favour of Type III approaches are as follows:

(i) if treatment effects vary from centre to centre then the only interpretable overall treatment effect would be a straightforward average of the centre effects: (ii) (a related point) if we use the treatment estimate as the basis of a test of hypothesis then the hypothesis we test concerns some average of the true treatment effects in each centre. It would be absurd if this hypothesized average depended on the numbers of patients we happened to have recruited to the trial.

The Type II proponent might argue as follows: (i) We really do not care which centres we recruit to the trial provided they deliver enough information. It is therefore nonsense not to weight them according to the amount of information they provide. (ii) If we consider all the centres we might have included but did not, then the Type III approach also depends arbitrarily on the numbers of patients recruited; it depends on whether the centre recruited none or some. (iii) Under the null hypothesis of no treatment effect there can be no treatment by centre interaction anyway and so any weighted combination of the treatment effects by centre forms a valid test of this hypothesis. Why not use the most efficient?

Senn also notes that even advocates of the Type III philosophy show Type II tendencies when they pool small centers. Senn supports the Type II analysis, he believes that the Type III philosophy is “untenable”(p. 1756). (Note: He briefly discusses the paradox of increased centers creating increased variance.)

Jones et al. (1998, 1767)) report the results of a simulation study “relating to the properties of various estimators of treatment differences in a ‘typical’ two-arm

parallel-group multi-centre clinical trial in hypertension.” Their results indicate that (p. 1776), “In terms of performance in estimating the overall average treatment difference, the Type III unweighted estimator performed poorly except when the variation in the treatment difference across centres was unrealistically large.” They also conclude that (p. 1777), “In terms of the power of various procedures to correctly identify a treatment difference, our study would suggest that the Type II weighted average approach has much to offer.”

Källén (1997, p. 932) also defends the Type II approach. He believes that “Taking the average over centers is an unnatural way of approaching the problem (of defining a treatment mean).” He continues (p. 935),

Average treatment effects over center means that mathematically, centers are studied instead of patients. It is argued that, in order to study patients, treatments should be analyzed by weighting center means according to center size, an old proposal but here brought forward as a natural approach based on ordinary least squares estimates.

Källén further argues that from a clinical perspective the weighted means are appropriate because the objective of a trial is to study the treatment effect on patients not centers.

Lin (1999, p. 370) points out what he terms to be a theoretical criticism of the Type II analysis: the weights are no longer interpretable in the presence of a treatment-by-center interaction and there is no well-defined population as a basis for the analysis. However, he uses the “Study Represented Population (SRP)”, to illustrate a solution to the problem. He believes that the weighted analysis is valid and interpretable, even in the presence of interaction.

Although there is still considerable discussion whether the most appropriate estimator of the overall treatment mean is the Type II or the Type III estimator, it is

clear that the variance of the Type II estimator is less than that of the Type III estimator. Hence it is a more efficient estimator.

Gallo (1998) shows that the variance of the unweighted estimator is always at least as large as the variance of the weighted estimator and that the two are equal only when the treatment-center sample sizes are identical. He notes that for the patterns of sample size imbalance that can be found in clinical trials, the variance of the Type III estimator can be substantially larger.

Jones et al. (1998) and Lin (1999) conducted simulation studies comparing the power of Type II and Type III models for various clinical trial scenarios. Their results confirm that the power advantage of the Type II model increases as sample size imbalance between centers increases.

An alternative criticism of the Type III model comes from Nelder (Nelder, J. A., 1994; Nelder, J. A. & Lane, P. W., 1995), who feels that Type III sums of squares are unnecessary constructs that arise from confusion concerning the linear model and antiquated methods needed before the advent of modern computers.

Regulatory guidance regarding the use of Type II and III models is evolving. As indicated above, the 1988 US Food and Drug Administration Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications (1988, p. 70) states, "Generally, it is recommended that SAS Type III or equivalent analyses be provided in addition to any other analyses." However, the 1998 International Conference on Harmonization Guidance on Statistical Principles for Clinical Trials (1998, p. 49589) states, "The main treatment effect may be investigated first using a model that allows for center differences, but does not include a term for treatment-by-center interaction." This would be a Type II model. The guideline indicates that if heterogeneity of treatment effects is found then (pp. 49589-

49590) “alternative estimates of the treatment effect, giving different weights to the centers, may be needed to substantiate the robustness of the estimates of the treatment effect.”

2.6 Analytical Studies Addressing the Analysis of Multicenter Trials

Three recent articles have reported on analytical studies addressing the analysis of data from multicenter trials. Gould (1998) used the analyses of data from a large multicenter trial to compare the results of empirical and conventional Bayes methods with the results of fixed and mixed model ANOVAs. Lin (1999) used simulated data to compare the power of using weighted and unweighted analyses with multicenter designs. Jones et al. (1998) reported on results of a simulation study of the analysis of data from a “typical” two-arm parallel-group multi-center clinical trial using four different estimators analyzed using a Bayesian approach and fixed and mixed model frequentist approaches.

Gould’s major interest in this article is the use of Bayes and empirical Bayes approaches; the fixed model ANOVA was the standard of the comparison. The fixed effect ANOVA model used by Gould was the conventional two-way factorial model with interaction. The treatment-by center interaction term was highly significant. Although, the author provided an interpretation of the source of the interaction, he did not discuss the appropriateness of the fixed effect model he used. The analysis strategies considered by Gould may be useful alternatives to the fixed effects ANOVA in some cases.

Lin simulated power calculation results for Type II and Type III analyses for a two-center design with unequal numbers of patients per treatment at the two centers. The Type II analysis has more power and the power advantage of the Type II analysis

increases as the sample imbalance increases. He also presents the results of a simulation comparing the powers for Type II and Type III analyses for 10, 20 and 30 centers with simulated patient numbers for the centers. These results confirm the power advantage of the Type II analysis. However, these simulations only treated the case when the treatment-by-center interaction is absent. Lin also discusses the appropriateness of using the Type II analysis in the presence of treatment-by-center interaction, but does not present any simulated results for that case.

Jones et al. simulated a “typical” trial comparing an active drug with a placebo in a parallel-group multi-center trial. Data were simulated using a model with fixed coefficients (SF), a model with random coefficients (SR) and a model with fixed interaction coefficients (SI). For each of the three simulation models, three or four different specifications were used to examine the effect of different relationships among the six centers. For model SF, the treatment difference for five of the centers was 10, whereas for the sixth center it was either = 5, 0, or -5, depending on the simulation run. For model SR, the overall between-treatment difference was 10, and the model coefficients were sampled from a normal distribution with mean zero and a range of standard deviations (SD): 2, 3.5, 5, and 10. The authors note that the SD of 10 is unrealistically large and was included to examine an extreme situation. The random error e_{ijk} was simulated from a normal distribution with mean zero and standard deviation 14. For model SI, the model coefficients were set to a fixed set of values that were obtained from a single sample from a normal distribution with standard deviation 2. Then in each simulation run the only sampling was over the distribution of the random error which had standard deviation of 14. To ensure an equivalent range in the treatment effect with the results from SR, the values of the

coefficients from SI were scaled by factors of 1, 1.75, 2.5, and 5 before being added to or subtracted from the treatment difference.

The simulated data were analyzed using three approaches: a fixed effects model (MF), a random effects model (MR), and a Bayesian model (MB). For the fixed effects model, three estimators of the treatment differences were calculated: the Type II estimate, the Type III estimate and a “pre-test estimate”. The pre-test estimate was determined using a two-stage testing procedure. The analysis model was first tested for significant interaction at a level of 10%. If the interaction was significant, the Type III estimator was used; if the interaction was not significant, the Type II estimator was used. The estimators were compared for each of the three simulated data sets using three criteria. The first was the MSE quantifying the difference between the estimated treatment effect and the parameter of interest. The second was the MSE measuring the accuracy of estimation achieved at each of the individual centers. The third was the power to detect a treatment difference.

The results of the comparisons with all three criteria indicate that in certain situations the Bayes or mixed model estimators are alternatives that may be preferable to the standard fixed effects model. The three fixed model approaches can only be compared using the overall treatment effect MSE and the power. For the individual treatment center MSE, all three fixed effects procedures gave the same results. For the overall treatment effect MSE, the Type II was generally superior to the Type III or Pre-test. The exceptions were the cases with large interaction, i.e. the case of the qualitative interaction in the SF simulation and the case with the “unrealistically” large SD in the SSR simulation (p. 1769). In both of those cases, the Type III approach was superior. When the power was used as the criteria, the Type II

approach was also superior in most cases. However, the Pre-test or Type III model was more powerful in case of large interaction.

Among the authors' conclusions are the following (pp. 1776-1777),

In this simulation study, a key problem has been that of defining a suitable overall measure of treatment efficacy when the treatment effect is not consistent across centres...

In terms of performance in estimating the overall average treatment difference, the Type III unweighted estimator performed poorly except when the variation in the treatment difference across centres was unrealistically large...

Pre-testing for treatment-by-centre interaction and then using a different estimator dependent on test outcome suffers from similar problems to pretesting for a carry-over difference in cross-over trials. The performance of the pre-testing approach in this simulation study provides little evidence to recommend it.

In terms of the power of various procedures to correctly identify a treatment difference, our study would suggest that the Type II weighted average approach has much to offer.

The use of simulation to compare rival estimators in multi-centre trials has proved quite difficult to implement, and the results need careful interpretation, nevertheless it provides some broad support for the increased use of the Type II estimators of overall treatment effects. This should be backed up by future analytical work.

The work of these authors demonstrates the utility of analytical studies in the solution of the issues related to the analysis of data from multicenter trials and underlines the need for additional work in this area.

2.7 Synthesis

There is considerable controversy associated with the statistical methodology for the analysis of data from multicenter trials. Statisticians disagree on the need to include the treatment-by-center interaction term in the primary analysis model. Some

propose that it should be an exploratory term. Many who propose that it should be included initially would conduct a preliminary test to determine if the term can be removed from the final analysis model. However, such a test may be underpowered and unreliable. The use of the Type II analysis is generally considered to be superior to the Type III analysis in the absence of treatment-by-center interaction. However, there is no consensus about the appropriate model in the presence of interaction. Some suggest that a Type II analysis is still appropriate while others feel that a Type III analysis must be used to eliminate bias. Several authors suggest that the more efficient Type II model could, as a minimum, be used when the interaction is quantitative, rather than qualitative, in nature.

The presence of quantitative interaction is felt by many authors to be inevitable and some suggest it should be viewed only as an exploratory factor. On the other hand, the presence of qualitative interaction may have a more important impact on the interpretation of any between-treatment differences. The ability to differentiate between qualitative and quantitative interaction would clarify the interpretation of the results in the presence of significant treatment-by-center interaction.

CHAPTER III

CURRENT METHODS FOR IDENTIFYING QUALITATIVE INTERACTION

3.1 Introduction

A fundamental premise in the conduct of a multicenter trial is the consistency across centers of the difference between the mean responses of patients in the respective treatment groups. This difference between treatments is sometimes called the “treatment effect”. Some variability in the treatment effect between centers is expected due to random variation, as well as demographic and clinical inconsistencies between centers. This between-center variability in the treatment effect is the so-called treatment-by-center interaction. As described in the previous chapter, the inclusion of this factor in the analysis model and its effect on the interpretation of the analysis is controversial. The analysis of data from multicenter trials must provide a method of differentiating random variability in the treatment effect from inherent between-center differences in the treatment effect.

The analysis of data from multicenter trials generally includes a test for the significance of the treatment-by-center interaction. When the presence of a treatment-by-center interaction is detected, one is often interested in investigating and explaining the factors that may contribute to the interaction. Also of concern is the classification of the interaction, i.e. determining whether the interaction is qualitative or quantitative. The presence of some quantitative interaction is not unexpected and, as Senn (1998) has shown, some degree of effect reversal may arise by chance alone when the number of centers is large. However, the presence of a sizeable interaction

demonstrating a large positive treatment effect at some centers and a large negative treatment effect at other centers is not usually expected and could have a serious impact on the further development of the tested treatment.

A common approach used to differentiate qualitative and quantitative interaction is to visually examine the means and the treatment differences across centers, either in a tabular or graphical presentation. If the mean treatment differences are not all positive or all negative, then the study is considered to show evidence of qualitative interaction. This method is practical, but does not provide any measure of confidence in the interpretation of the results.

Ciminera, Heyse, Nguyen and Tukey (1993), Gail and Simon (1985) and Azzalini and Cox (1984) have each proposed statistical methods that are applicable to the problem of identifying the presence of qualitative interaction for multicenter trials. Each method also provides a measure of the evidence supporting the existence of a qualitative interaction, although the approaches and the theory of the three methods differ. We will describe each method. However, we will first define some common terms, which will be used in the presentation of each method, and present an analysis of variance model for a multicenter trial.

3.2 Definition of Terminology

Let there be a multicenter trial conducted with b treatment levels, $i = 1$ to b , and at c centers, $j = 1$ to c . And let the measure of the response of the patients receiving treatment i at center j be μ_{ij} . We may be interested in the difference, δ , between two treatments groups, say i' and i'' . The difference between the responses of the two treatment groups at center j is

$$(3.1) \quad \delta_j = \mu_{i'j} - \mu_{i''j}.$$

(Note: Since δ_j is the difference between the responses of two designated treatments, i' and i'' , at center j , δ_j could be labeled more precisely as $\delta_{(i', i'', j)}$. In this study, we will never simultaneously examine differences between more than two treatments. Hence, for simplicity of notation, we will use δ_j to denote the difference between two designated treatment responses at center j .)

Suppose that we have two treatment levels, $i = i', i''$ and two centers, $j = j', j''$. Let the four cell responses be $\mu_{i'j'}$, $\mu_{i''j'}$, $\mu_{i'j''}$, $\mu_{i''j''}$. A treatment-by-center interaction exists within these four cells if the difference in response for the two treatment groups differs between the two centers, that is

$$(3.2) \quad \mu_{i'j'} - \mu_{i''j'} \neq \mu_{i'j''} - \mu_{i''j''}.$$

3.2.1 Quantitative Interaction

When the differences vary in magnitude, but not in direction of treatment effects between centers (i.e. the pairs of responses have the same sign), then the interaction is a quantitative interaction. If a quantitative interaction exists, then the condition in (3.2) exists and

$$(3.3) \quad \mu_{i'j'} - \mu_{i''j'} \geq 0 \text{ and } \mu_{i'j''} - \mu_{i''j''} \geq 0, \text{ or}$$

$$(3.4) \quad \mu_{i'j'} - \mu_{i''j'} \leq 0 \text{ and } \mu_{i'j''} - \mu_{i''j''} \leq 0.$$

3.2.2 Qualitative Interaction

However, if the differences between the pairs of responses do not have the same sign, then the interaction is a qualitative interaction. If a qualitative interaction exists, then the condition in (3.2) exists and

$$(3.5) \quad \mu_{i'j'} - \mu_{i''j'} > 0 \text{ and } \mu_{i'j''} - \mu_{i''j''} < 0, \text{ or}$$

$$(3.6) \quad \mu_{i'j'} - \mu_{i''j'} < 0 \text{ and } \mu_{i'j''} - \mu_{i''j''} > 0.$$

In terms of δ_j , the definitions of interaction [(3.2) - (3.6)] are restated below.

- Interaction if $\delta_j \neq \delta_{j''}$
- Quantitative interaction if $\delta_j \neq \delta_{j''}$ and $[(\delta_j \geq 0 \text{ and } \delta_{j''} \geq 0) \text{ or } (\delta_j \leq 0 \text{ and } \delta_{j''} \leq 0)]$
- Qualitative interaction if $\delta_j \neq \delta_{j''}$ and $[(\delta_j > 0 \text{ and } \delta_{j''} < 0) \text{ or } (\delta_j < 0 \text{ and } \delta_{j''} > 0)]$.

These three conditions are represented graphically in Figures 1 – 3.

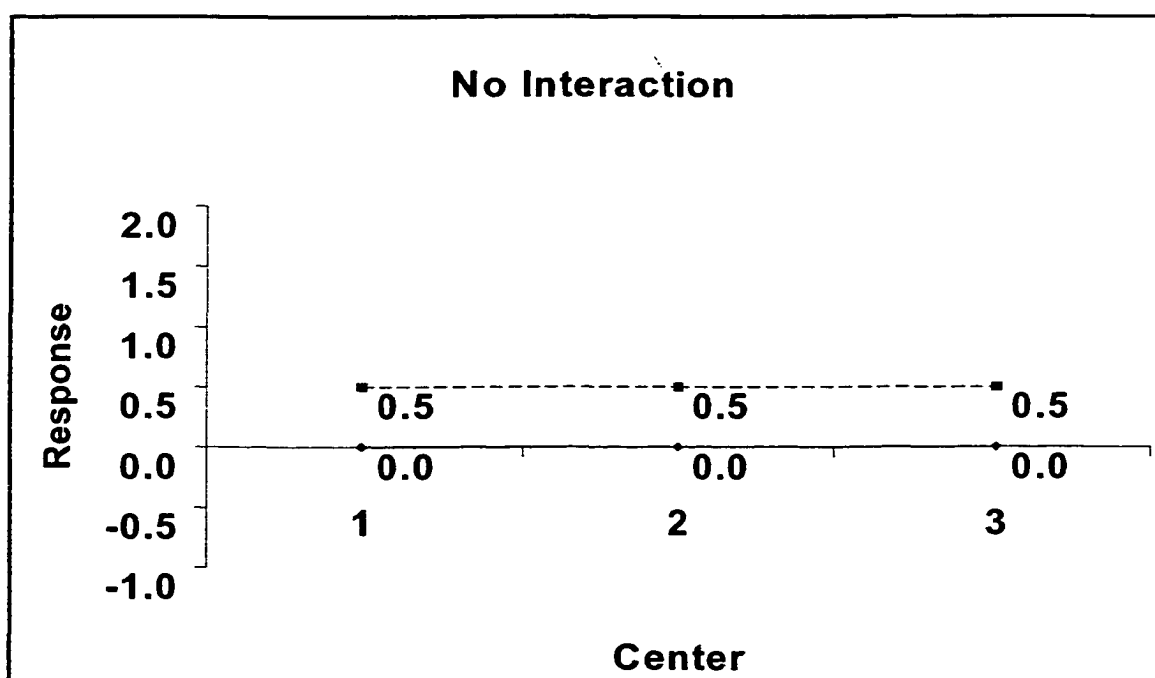


Figure 1. Mean Responses for Two Treatments (Treatment 1 —, Treatment 2 --) at Three Centers with No Interaction.

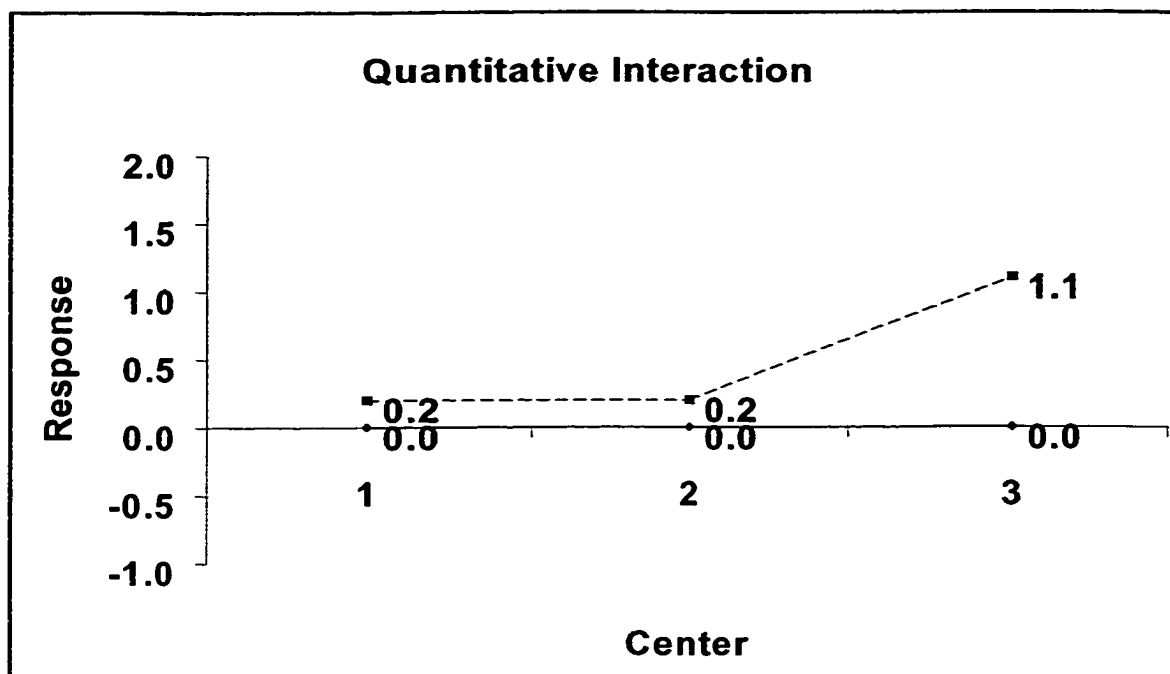


Figure 2. Mean Responses for Two Treatments (Treatment 1 —, Treatment 2 ---) at Three Centers with Quantitative Interaction.

3.3 Estimation of Parameters

The definitions above are given in terms of the population parameters.

However, in the three methods referenced above, inferences are generally based upon sample estimates of these parameters.

The sample estimate of the response of the patients receiving treatment i at center j will be termed $\hat{\mu}_{ij}$. The determination of this estimator may vary among the three methods and will be further defined in the descriptions of the methods. The estimates of the variance of $\hat{\mu}_{ij}$ may also vary among the different methods; they will not be defined here, but will be discussed separately for each method.

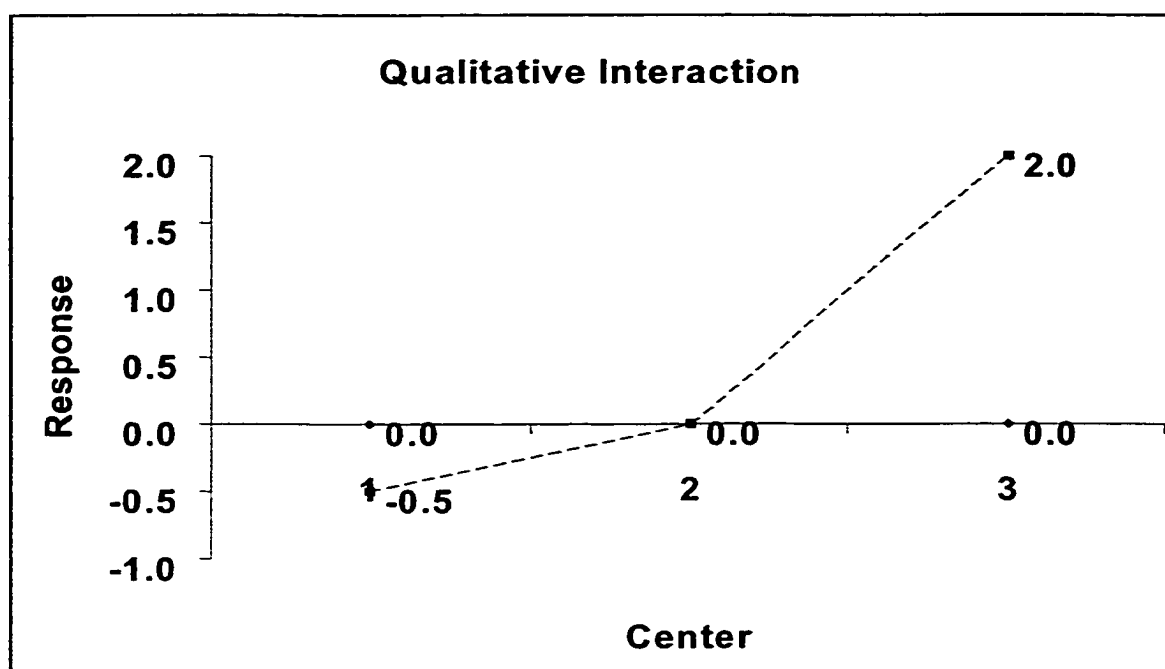


Figure 3. Mean Responses for Two Treatments (Treatment 1 —, Treatment 2 --) at Three Centers with Qualitative Interaction.

The sample estimate of the difference in the responses of two treatment groups (i' and i'') at center j , (i.e. the estimate of δ_j) regardless of the measure of response used, is

$$(3.7) \quad d_j = \hat{\mu}_{i'j} - \hat{\mu}_{i''j}.$$

(Note: Since d_j is the sample estimate of the difference between the responses of two designated treatments, i' and i'' , at center j , d_j could be labeled more precisely as $d_{(i', i''); j}$. In this study, we will never simultaneously examine differences between more than two treatments. Hence, for simplicity of notation, we will use d_j to denote the difference between two designated treatment responses at center j .)

The estimate of the variance of the difference in the responses of two treatment groups at center j will be discussed in the description of each method.

3.4 Specification of the Analysis of Variance Model

Continuous response data from a multicenter trial can often be described with the following model:

$$(3.8) \quad y_{ijk} = \mu + B_i + C_j + (BC)_{ij} + e_{ijk}, \text{ where}$$

$i = 1, \dots, b$ (number of treatments),

$j = 1, \dots, c$ (number of centers),

$k = 1, \dots, n_{ij}$ (number of patients assigned to treatment i at center j),

y_{ijk} is the measured response of patient k assigned to treatment i at center j ,

μ is the overall mean response,

B_i is the fixed effect of the i th treatment (the treatment mean of treatment i is $\mu_i = \mu + B_i$),

C_j is the fixed effect of the j th center,

$(BC)_{ij}$ is the fixed interaction effect of the i th treatment and the j th center and

e_{ijk} is the random error associated with patient k assigned to treatment i at center j . The e_{ijk} are assumed to be independent, $N(0, \sigma^2)$.

Furthermore, the mean response of the patients receiving treatment i at center j is $\mu_{ij} = \mu + B_i + C_j + (BC)_{ij}$. Each sample of patients receiving any treatment, i , at any center, j , is assumed to be independent, $N(\mu_{ij} = \mu + B_i + C_j + (BC)_{ij}, \sigma^2)$. Note that all of these samples are assumed to have a common variance, $\sigma^2_{ij} = \sigma^2$.

The parameters of the model can be estimated and tested using analysis of variance. The primary interest in the analysis is the test of significance of the different levels of the treatment effect, B_i , i.e. the test of significance of the treatment's influence on the response. Also of interest is the test of significance of the levels of $(BC)_{ij}$, the test of the treatment-by-center interaction.

The traditional analysis of variance tests the hypotheses using mean squares. The mean squares associated with the treatment effect (MSB) and the treatment-by-center interaction (MSBC) are both tested for significance with F-tests based on the

error mean square (MSE), which is an estimate of σ^2 . This approach is illustrated in the standard analysis of variance table presented below in Table 1

Table 1
Analysis of Variance Table for a Multicenter Trial

Source	Degrees of freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F Value
Treatment (B)	dfB = b-1	SSB	MSB = SSB / dfB	MSB / MSE
Center (C)	dfC = c-1	SSC	MSC = SSC / dfC	MSC / MSE
Treatment X Center (BC)	dfBC = (b-1)(c-1)	SSBC	MSBC = SSBC / dfBC	MSBC / MSE
Error (E)	dfE = $\sum (n_{ij} - 1)$	SSE	MSE = SSE / dfE	

For this model, a sample estimate of the response of the patients receiving treatment i at center j is the cell mean

$$(3.9) \quad \hat{\mu}_{ij} = \bar{y}_{ij} = (\sum_k y_{ijk}) / n_{ij}.$$

We will call the estimate of σ^2 from analysis of variance s_A^2 ,

$$(3.10) \quad s_A^2 = \text{estimate of } \sigma^2 \text{ from analysis of variance (corresponds to MSE)}.$$

We can use this estimate of the sample variance to calculate the estimated variance of any $\hat{\mu}_{ij} = \bar{y}_{ij}$. Although the variance is assumed to be constant across all

centers and treatment groups, the variance of each \bar{y}_{ij} is also a function of the sample size, n_{ij} .

When the sample sizes, n_{ij} , differ across both treatment groups and centers, it follows that the means for different treatment-by-center cells (\bar{y}_{ij}) have different variances, dependent upon their sample size,

$$(3.11) \quad s_{A\bar{y}_{(ij)}}^2 = s_A^2 / n_{ij}.$$

The variance of the difference between the means of two treatment groups, $i = i', i''$, within a given center j is,

$$(3.12) \quad s_{A d(j)}^2 = s_A^2 [(1 / n_{i'j}) + (1 / n_{i''j})].$$

(Note: Since $s_{A d(j)}^2$ is a sample estimate of the variance of the difference between the responses of two designated treatments, i' and i'' , at center j , $s_{A d(j)}^2$ could be labeled more precisely as $s_{A d(i', i'', j)}^2$. In this study, we will never simultaneously examine differences between more than two treatments. Hence, for simplicity of notation, we will use $s_{A d(j)}^2$ to denote the variance of the difference between two designated treatment responses at center j .)

For the special case where the sample sizes differ across centers, but are equal for the treatment groups within a center j , $n_j = n_{i'j} = n_{i''j}$, (3.12) can be simplified as,

$$(3.13) \quad s_{A d(j)}^2 = 2s_A^2 / n_j.$$

For the case where the n_{ij} are equal across all treatment groups at all centers (i.e. $n_{ij} = n$), the variance of any \bar{y}_{ij} is

$$(3.14) \quad s_{A\bar{y}}^2 = s_A^2 / n.$$

And the estimate of the variance of the difference between the means of any two treatment groups at any given center is

$$(3.15) \quad s_{A_d}^2 = 2 s_A^2 / n.$$

3.5 Ciminera, Heyse, Nguyen and Tukey

Ciminera and his co-authors (1993) present a method that is not a formal test for the presence of qualitative interaction, but a method of validating the presence of “substantial evidence” of qualitative interaction. Their method is based on the signs (positive or negative) of the differences in the responses of two treatment groups across study centers. They maintain that (p. 1034), “Merely the appearance of reversed means (i.e. differences with differing signs) at some centres is, thus, not in itself substantial evidence of qualitative interaction.”

Furthermore, they state that (p. 1034),

If we are to have an appropriate basis for deciding whether treatment-by-centre interaction should contribute to our estimated uncertainty of the overall trial result, we must have an analysis which makes allowance for the uncertainties in the individual centre results and decides whether the few reversed centre means found are to be treated as mere consequences of sampling variation, inestimable and estimable...or whether the reversed centre means are to be treated as indicating reversed many-patient averages...

They recommend what they feel to be an appropriate basis.

If we knew that the variances of measurement were all equal, and if we knew this common value, we could look at the most negative observed value for any centre and ask whether its deviation from zero, conveniently expressed as a fraction or multiple of its standard deviation, is large enough for it to be unlikely to correspond to a zero or positive many-patient average effect. This is a question about an order statistic, the most negative of the observed values, and we can reasonably compare the observed negative deviation with the mean or median of the distribution of the lowest order statistic from a distribution with known variance and mean zero. If the observation is

less negative, it yields no evidence of a real negative effect.

A convenient way to make the comparison is by first 'pushing back' the extreme value, shifting it by the mean or median of the appropriate order statistic distribution and then asking if the result is still negative.

Building upon this method of "pushing back", the authors develop a method to determine if there is "substantial evidence" of qualitative interaction. They first relax the assumption of a common estimate of the variance of centre means (p. 1034).

A natural and effective way, among others, to make allowance for the differing estimated variances of the centre means is to choose a reference value (near the middle of the ordered array of centre means) and to convert the individual centre results into standardized deviations (standardized using the standard error for that centre's results, estimated internally at each centre) from this reference value. The deviations can then be ordered, and the results 'pushed back' by amounts appropriate for the corresponding order statistics in a sample from the relevant distribution.

They recommend three modifications of the 'pushback' process to avoid excessive conservatism (pp. 1035-1036).

First, we can use Student's t-distribution to obtain amounts of pushback for the standardized deviations. Second, we can seek (and use in place of order statistic medians) outward per cent points of the corresponding order-statistic distributions (for example, the lower 10 per cent point for centres below the median and the upper 90 per cent point for those above the median), rather than the medians, since actual random sampling fluctuations will often correspond to quite extreme location in its order-statistic distribution rather than to the median. Thirdly, we can group and pool within groups the internally estimated standard deviations of the centre means.

Although they do not explicitly discuss assumptions on the data that are necessary for their method to be valid, they present their method in the context of the analysis of variance model presented above, with center as a fixed factor. The authors implicitly define the sample estimate of the response of the patients receiving treatment i at center j ($\hat{\mu}_{ij}$) as the mean response, \bar{y}_{ij} (3.9). However, they do not use

the analysis of variance estimate of σ^2 , i.e. s_A^2 . Their recommended variance estimation procedure, which (p. 1034) “make(s) allowance for the differing estimated variances of the centre means” will be presented below.

Their method evaluates qualitative interaction between two treatment groups at multiple centers. However, they do not extend their method to the case of more than two treatment groups.

The authors present several options for the implementation of the method and review some possible choices for a measure of overall central tendency, the estimate of overall variability, the appropriate distribution, and the appropriate amount of pushback. Their recommendation is to use the least conservative combination. The implementation of the method they recommend is described hereafter.

The sample estimates of the mean response of the patients receiving each treatment i at each center j (3.9) are calculated and used to determine the sample estimates of the difference in the mean responses of the two treatment groups at each center j (3.7).

Ciminera et al. estimate the variance of $\hat{\mu}_{ij} = \bar{y}_{ij}$ internally within each treatment group at each center. This estimate of the variance of each mean, as defined by the authors, will be termed $s_{\bar{y}_{ij}}^2$,

$$(3.16) \quad s_{\bar{y}_{ij}}^2 = s_{\bar{y}_{ij}}^2 / n_{ij}, \text{ where}$$

$$(3.17) \quad s_{\bar{y}_{ij}}^2 = \sum_k (y_{ijk} - \bar{y}_{ij})^2 / (n_{ij} - 1).$$

The estimated variance of the difference in the mean responses of two treatment groups, $i = 1, 2$, at center j is obtained by first pooling the estimates of the variances of the samples from which the two means were obtained.

$$(3.18) \quad s_{P(d)}^2 = [(n_{1j} - 1)s_{1j}^2 + (n_{2j} - 1)s_{2j}^2] / (n_{1j} + n_{2j} - 2).$$

It follows that the estimate of the variance of the difference in the mean responses of the two treatment groups at center j is

$$(3.19) \quad s_{P(d)}^2 = s_{P(d)}^2 [(1 / n_{1j}) + (1 / n_{2j})].$$

A pooled overall estimate across centers of the between treatment variance for the two designated treatments is calculated as described below (3.20)

$$(3.20) \quad s_{P(overall)}^2 = [(\sum_j v_{d(j)} s_{P(d)}^2) / \sum_j v_{d(j)}], \text{ where}$$

$$v_{d(j)} = n_{1j} + n_{2j} - 2.$$

(Note: Since $s_{P(d)}^2$ is a sample estimate of the variance of the difference between the responses of two designated treatments, 1 and 2, at center j , $s_{P(d)}^2$ could be labeled more precisely as $s_{P(d(1,2;j))}^2$. It follows that $s_{P(overall)}^2$ could be more precisely labeled as $s_{P(overall(1,2))}^2$. In this study, we will never simultaneously examine differences between more than two treatments. Hence, for simplicity of notation, we will use $s_{P(d)}^2$ to denote the variance of the difference between two designated treatment responses at center j and $s_{P(overall)}^2$ to denote a pooled overall estimate across centers of the between treatment variance for the two designated treatments.)

The sample estimates of the difference in the mean responses of the two treatment groups at each center (d_j) are used to determine a weighted median difference across centers (m). The median (m) is calculated by taking the median of a vector of mean treatment differences, constructed by including the mean treatment difference at each center once for each patient in the trial at the center. For example, suppose the mean difference at center 3 between treatment 1 ($n_{13}=14$) and treatment 2 ($n_{23}=12$) is 2.0, the vector of mean values would include 26 observations with the

value of 2.0. The median of this weighted vector of mean differences is the overall median difference (m) used for this method.

The median difference (m) is subtracted from the mean treatment difference from each center (d_j) to determine a deviation for each center. Each deviation is then standardized by dividing by the pooled standard deviation from (3.20).

$$(3.21) \quad D_{s(j)} = (d_j - m) / SP_{\text{overall}}$$

These standardized deviations ($D_{s(j)}$) are then ordered by magnitude and “pushed back” by amounts appropriate for the order statistic corresponding to the rank order of the d_j for the center, as described below.

The d_j of the centers are ordered and the centers are ranked in ascending order according the magnitude of the corresponding d_j . For centers with d_j equal to or below the overall median difference (m), the lower 10 per cent point of the incomplete beta function is calculated using parameters: v_1 = the rank order of the center and v_2 = (total number of centers – rank order of the center) + 1.

$$(3.22) \quad P_j = 10\text{th percentile of } \beta(v_1, v_2), \text{ for all } d_j \leq m.$$

For centers with d_j above the median (m), the upper 10 per cent point (i.e. the 90 per cent point) of the incomplete beta function is calculated using parameters: v_1 and v_2 , as defined above.

$$(3.23) \quad P_j = 90\text{th percentile of } \beta(v_1, v_2), \text{ for all } d_j > m.$$

The percent points of the incomplete beta function calculated for each center are then used as percentiles of Student’s t distribution and the t value (t_j) corresponding to each value is determined. For each center, t_j is the value that gives a

cumulative cdf for Student's t-distribution, with the respective df, equal to the corresponding P_j calculated above ((3.22) or (3.23)),

$$(3.24) \quad t_j = t(P_j, (n_{1j} + n_{2j} - 2)).$$

The t_j value calculated for each center is then subtracted from the standardized deviation for the center,

$$(3.25) \quad D_{p(j)} = D_{s(j)} - t_j.$$

However, if the sign of $D_{p(j)}$ is opposite to that of $D_{s(j)}$, the result is zero (to prevent pushback across zero). Zero is always pushed-back to zero, i.e. if $D_{s(j)} = 0$, then $D_{p(j)} = 0$.

The pushed-back standardized deviations are restored to their original scale (destandardized) by multiplying by the pooled standard deviation (from (3.20)) and adding the overall median difference (m),

$$(3.26) \quad D_{d(j)} = (S_P \text{ overall} * D_{p(j)}) + m.$$

The authors conclude (p. 1035) that the “appearance of opposite signs among the destandardized values ($D_{d(j)}$) may be taken as ‘substantial evidence’ of qualitative interaction”. That is, if there is no evidence of qualitative interaction, then the signs of the destandardized values ($D_{d(j)}$) should be either all positive or all negative. The presence of at least one center with a positive $D_{d(j)}$ in conjunction with at least one center with a negative $D_{d(j)}$ is “substantial evidence” of qualitative interaction.

Although the procedures presented here are those prescribed by Ciminera et al.; using other procedures outlined by the authors may be advisable in certain situations. For example, when there is evidence that the variance of the difference in

the responses of two treatment groups is heterogeneous across centers, the authors recommend pooling across centers with homogeneous variance the internal estimates of the variance of the between-treatment differences. The authors also note that the method is applicable to test for qualitative interaction between treatment groups and other subgroups of patients besides centers.

3.6 Gail and Simon

Gail and Simon (1985) assume that two treatments are being compared and that the true treatment differences between these two treatments for each subgroup of patients, δ_j , have estimates which are independent and normally distributed with mean δ_j and known variance $\sigma^2_{d(j)}$. These estimates of δ_j , termed d_j , are previously defined in (3.7). The authors note that the procedures they introduce are valid for large samples if consistent estimates of $\sigma^2_{d(j)}$ are used instead. They also indicate that their method works for independent d_j with possibly unequal variances. Their method is designed for independent estimates of d_j ; but is also applicable if the d_j are estimated using a model fitted to the entire dataset, if the estimates of the d_j are uncorrelated. (Note: As previously designated, for simplicity of notation, we will use d_j to denote the difference between two designated treatment responses at center j and $\sigma^2_{d(j)}$ to denote the variance of d_j . A more precise notation would be $d_{(i', i'', j)}$ and $\sigma^2_{d(i', i'', j)}$.)

The authors present their method in a context more general than that of the analysis of variance model described above and illustrate the method with an example using binomial data. Gail and Simon state (p. 364) that “the asymptotic normality of most estimators of relative treatment efficacy renders this test broadly applicable.” We also remark that the method is applicable to subgroups of patients other than centers.

Their method evaluates qualitative interaction between two treatment groups across multiple subgroups. However, they do not extend their method to the case of more than two treatments.

The terminology used by Gail and Simon to define the types of interaction differs from the terminology in standard usage. They define three types of interactions: (1) quantitative, (2) non-crossover, and (3) qualitative or crossover. Their definition of quantitative interaction (p. 361), “any heterogeneity of treatment effects among subsets of patients”, corresponds to (3.2) and is commonly used in a more general sense to define any interaction, with quantitative and qualitative being subsets. Gail and Simon label the type of interaction that is generally termed “quantitative” as “non-crossover” (i.e.(3.3) or (3.4)). Their definition of qualitative interaction (i.e. (3.5) or (3.6)) is consistent with that commonly used; they also call this type of interaction “crossover interaction”. For consistency, the descriptions of their methods will use the common terms applied elsewhere, i.e. (1) interaction, (2) quantitative, and (3) qualitative.

The first test presented by the authors is a test of the hypothesis of no qualitative interaction. This is equivalent to the hypothesis that the vector of treatment differences $\Delta = (\delta_1, \dots, \delta_c)$, for subgroups $j = 1$ to c , lies either in the region in which all components are nonnegative (call this region \mathbf{O}^+) or in the region in which all components are nonpositive (call this region \mathbf{O}^-). That is, $\mathbf{O}^+ = \{\Delta: \delta_j \geq 0 \text{ all } j\}$ and $\mathbf{O}^- = \{\Delta: \delta_j \leq 0 \text{ all } j\}$. The likelihood ratio test is

$$(3.27) \quad \max_{\Delta \in \mathbf{O}^+ \cup \mathbf{O}^-} \exp \sum_j [-(d_j - \delta_j)^2 / (2\sigma_{d(j)}^2)] < s,$$

where the constant s is chosen to ensure that the rejection region does not exceed significance level α for any point in the null space $\mathbf{O}^- \cup \mathbf{O}^+$.

The inequality (3.27) is equivalent to two simultaneous inequalities

$$(3.28) \min_{\Delta \in \mathbf{O}^-} \sum_j (d_j - \delta_j)^2 / \sigma_{d(j)}^2 > k, \text{ and}$$

$$(3.29) \min_{\Delta \in \mathbf{O}^+} \sum_j (d_j - \delta_j)^2 / \sigma_{d(j)}^2 > k,$$

where $k = -2 \log(s)$.

Thus the null hypothesis can be rejected when the d_j are far away from both \mathbf{O}^+ and \mathbf{O}^- , with distance defined by the inverse variance metric. The minimum value for (3.28) occurs for $\delta_j = d_j$ if $d_j = 0$ and for $\delta_j = 0$ otherwise. Similarly, the minimum value for (3.29) occurs for $\delta_j = d_j$ if $d_j = 0$ and for $\delta_j = 0$ otherwise. Hence, we reject the hypothesis of no qualitative interaction if both

$$(3.30) Q^- = \sum_j (d_j^2 / \sigma_{d(j)}^2) I(d_j > 0) > k, \text{ and}$$

$$(3.31) Q^+ = \sum_j (d_j^2 / \sigma_{d(j)}^2) I(d_j < 0) > k, \text{ where}$$

$I(d_j > 0) = 1$ if $d_j > 0$ and 0 otherwise, and

$I(d_j < 0) = 1$ if $d_j < 0$ and 0 otherwise.

The quantities Q^+ and Q^- are minimum values of $\sum_j (d_j - \delta_j)^2 / \sigma_{d(j)}^2$ over \mathbf{O}^+ and \mathbf{O}^- , respectively, and the likelihood ratio test can be expressed as $\min(Q^+, Q^-) > k$. Gail and Simon show that the distribution of k is the sum of a series of conditional probabilities, which can be calculated as the sum of the products of binomial mass functions and corresponding chi-squares. This distribution can be used to determine the level of significance of a calculated value of k and to establish values of k such that for all $\Delta \in \mathbf{O}^+ \cup \mathbf{O}^-$, the probability that (3.30) and (3.31) are both satisfied is no greater than the significance level, α .

To determine the significance level of the test for qualitative interaction using this method we first let,

$$(3.32) k = \min(Q^+, Q^-).$$

We then use the distribution of k to determine the level of the test,

$$(3.33) \quad \alpha = \sum_{j=1}^{c-1} B(j; n = c - 1, p = 0.5) [1 - F_j(k)], \text{ where}$$

$B(j; n, p)$ is the binomial probability mass function with index n and parameter p ,
 $F_j(k)$ is the central chi-square distribution with j degrees of freedom,
 and
 c is the number of subgroups of patients.

If the value from (3.33) is less than the predetermined α , then the null hypothesis of no qualitative interaction is rejected.

The authors also present a test of the null hypothesis that a designated treatment is at least as good as the other treatment in every subgroup of patients. This null hypothesis of non-inferiority can be expressed as $\Delta \in O^+$. It follows that the likelihood ratio test is $Q^+ > k^*$. Gail and Simon determine the distribution of k^* , where

$$(3.34) \quad k^* = Q^+.$$

They show that the distribution of k they present can also be used to determine the level of the test of this null hypothesis. The appropriate significance level can be obtained by substitution of c for $c-1$ in the equation presented above (3.33),

$$(3.35) \quad \alpha = \sum_{j=1}^c B(j; n = c, p = 0.5) [1 - F_j(k^*)], \text{ where}$$

$B(j; n, p)$ is the binomial probability mass function with index n and parameter p ,
 $F_j(k^*)$ is the central chi-square distribution with j degrees of freedom,
 and
 c is the number of subgroups of patients.

If the value from (3.35) is less than the predetermined α , then the null hypothesis of non-inferiority is rejected.

Similarly, the null hypothesis $\Delta \in \mathbf{O}^*$ can be tested with $Q^* > k^*$, which is accomplished with a substitution of Q^* into (3.34).

A third test presented by Gail and Simon is the test of no interaction, based on the weighted residual sum of squares,

$$(3.36) \quad H = \sum_j (d_j - \bar{d})^2 / \sigma_{d(j)}^2, \text{ where}$$

$$(3.37) \quad \bar{d} = [\sum_j (d_j / \sigma_{d(j)}^2)] / [\sum_j (1 / \sigma_{d(j)}^2)] \text{ is a weighted mean.}$$

Under the hypothesis of no quantitative interaction, H has a central chi-square distribution with degrees of freedom, $df = c - 1$, where c is the number of subgroups.

Gail and Simon note that consistent estimates of $\sigma_{d(j)}^2$ may be inserted in all of the equations presented above without altering the asymptotic distribution theory.

3.7 Azzalini and Cox

The underlying assumptions specified for the test proposed by Azzalini and Cox (1984) are more extensive, and restrictive, than those of the other methods. Azzalini and Cox assume the normal-theory linear model, with an error variance that is known. A known error variance implies (p. 335) “in effect that the number of degrees of freedom for estimating error is large”. The authors note that the effect of using an estimate of σ^2 has been briefly investigated by simulation. They conclude (p. 339) that, “If the degrees of freedom for estimating σ are at least 30, there is little effect on the significance levels. If the degrees of freedom are between 10 and 30, an apparent 5 per cent level is at worst 10 per cent and an apparent 1 per cent level is at worst 3 per cent.” Hence if σ is estimated from a sample size of less than 30, then the probability of detecting a qualitative interaction could be somewhat higher than the nominal level. Consistent with their assumptions, they use the mean square within

cells (MSE) as an estimator of the error variance in their example. We have called this estimator s_A^2 (3.10). Also implicit in their method is the definition of the sample estimate of the response of the patients receiving treatment i at center j as the cell mean, i.e. $\hat{\mu}_{ij} = \bar{y}_{ij}$ (3.9).

The test they propose and the distributional calculations for the test statistic are based on the global null hypothesis of no treatment effects. They show that the test is conservative in the presence of main effects.

The authors do not mention any limit on the number of treatment groups that can be simultaneously tested for qualitative interaction with their test. Thus, in contrast to the two previously discussed methods, this method allows simultaneous evaluation of qualitative interaction among multiple treatment groups at multiple centers.

An important, and restrictive, assumption for this test is that of equal sample sizes across treatments and centers, i.e. $n_{ij} = n$. This condition implies that for all \bar{y}_{ij} , $s_{A\bar{y}}^2 = s_A^2 / n$ (3.14), and that the estimate of the variance of the difference in the mean responses of any two treatment groups is the same at all centers, i.e. $s_{A_d}^2 = 2 s_A^2 / n$ (3.15).

The basis of this test for qualitative interaction is the examination of the patterns of differences in quadruples of mean responses, i.e. the mean responses for two treatment groups at two centers. The authors motivate their test based on the condition for qualitative interaction presented in (3.5). (Note that for two centers j' and j'' , the conditions in (3.5) and (3.6) are interchangeable depending on which center is defined as j' and which is j'' .)

Suppose that we have two treatment levels, $i = i', i''$ and two centers, $j = j', j''$.

Let the four true cell means be $\mu_{i'j'}, \mu_{i''j'}, \mu_{i'j''}, \mu_{i''j''}$. A qualitative interaction exists within these four cells if the condition in (3.5) exists, that is

$$(3.38) \quad \mu_{i'j'} - \mu_{i''j'} > 0 \text{ and } \mu_{i'j''} - \mu_{i''j''} < 0.$$

Suppose that the estimates of these cells means, $\bar{y}_{i'j'}, \bar{y}_{i''j'}, \bar{y}_{i'j''}, \bar{y}_{i''j''}$, are based on samples of equal size, n , and have a common, known variance, σ^2 / n , then the standard error of a simple contrast estimating the condition of qualitative interaction (3.38) would be $\sigma\sqrt{(2/n)}$. We note that this is the square root of the variance of the difference between two means, with equal sample sizes, n , i.e.,

$$(3.39) \quad \sigma_d^2 = 2\sigma^2 / n.$$

Evidence for a qualitative interaction would be provided if, for a sufficiently large positive value, say t ,

$$(3.40) \quad \bar{y}_{i'j'} - \bar{y}_{i''j'} \geq t\sigma_d \text{ and } \bar{y}_{i'j''} - \bar{y}_{i''j''} \leq -t\sigma_d.$$

We can rewrite this expression in terms of $d_j = \bar{y}_{i'j} - \bar{y}_{i''j}$ (3.7) as,

$$(3.41) \quad d_{j'} \geq t\sigma_d \text{ and } d_{j''} \leq -t\sigma_d.$$

More generally, suppose that we have b treatment levels, $i = 1$ to b , and c centers, $j = 1$ to c . Azzalini and Cox propose as a test statistic, say t^* , the largest t such that for some quadruple of cells with sample means, $\bar{y}_{i'j'}, \bar{y}_{i''j'}, \bar{y}_{i'j''}$ and $\bar{y}_{i''j''}$, the relation in (3.41) is satisfied.

Then t^* is the largest value such that among all pairs of within center treatment differences, d_j , there exist two differences which meet the condition of (3.41). That is,

$$(3.42) \quad d_j \geq t^* \sigma_d \text{ and } d_{j^*} \leq -t^* \sigma_d.$$

If t^* is small, then among all the d_j relatively large negative and positive differences do not exist. A larger t^* is indicative of the presence of relatively large negative and positive differences. The significance level of t^* can be calculated to determine the presence of significant qualitative interaction.

Let S be the total number of quadruples of cells satisfying (3.42). We need to find $P(S=0)$, or alternatively we must find t^* such that $P(S=0) = 1 - \alpha$, for given α . As t^* increases, $E(S) \rightarrow 0$ and Azzalini and Cox show that, provided that $P(S>1)$ is negligible compared to $P(S=1)$, that it is reasonable to anticipate that $1 - P(S=0) \sim E(S)$. They show, based on a lemma provided in their paper, that a Poisson approximation of the distribution of S exists.

$$(3.43) \quad P(S=0) \sim \exp [-0.5c(c-1)b(b-1) \{\Phi(-t^*)\}^2], \text{ where}$$

c is the number of centers,

b is the number of treatment levels, and

$\Phi(t^*)$ is the cdf from the standard unit normal distribution of (t^*) .

We can calculate t^* using the estimator of σ_d based on the analysis of variance (3.15). It follows that (3.42) becomes,

$$(3.44) \quad d_j \geq t^* s_{Ad} \text{ and } d_{j^*} \leq -t^* s_{Ad}.$$

For each pair of differences where the sign of $d_{j'}$ differs from that of $d_{j''}$, the absolute values of the two differences are compared and the smaller of the two absolute values, say $d^*_{j'j''}$, is selected,

$$(3.45) \quad d^*_{j'j''} = \min (|d_{j'}|, |d_{j''}|),$$

where the sign of $d_{j'}$ differs from that of $d_{j''}$.

The smaller absolute differences from all such pairs ($d^*_{j'j''}$) are then compared. The largest of all these smaller absolute differences is determined,

$$(3.46) \quad w = \max (d^*_{j'j''}).$$

This maximum difference (w) is used in the calculation of the test statistic. The test statistic is

$$(3.47) \quad t^* = w / s_{Ad}.$$

The level of significance of t^* is

$$(3.48) \quad 1 - \exp [-0.5c(c-1)b(b-1) \{\Phi(-t^*)\}^2], \text{ where}$$

c is the number of centers,

b is the number of treatment levels, and

$\Phi(t^*)$ is the cdf from the standard unit normal distribution of (t^*).

If the value from (3.48) is less than the predetermined α , then the null hypothesis of no qualitative interaction is rejected and the presence of at least one quadruple with $\mu_{i'j'} - \mu_{i''j'} > 0$ and $\mu_{i'j''} - \mu_{i''j''} < 0$ is confirmed.

3.8 Evaluation and Comparison of Methods -- Previous Work

The authors of all three papers present numerical examples illustrating their methods. However, discussion of the operating characteristics of the proposed methods is minimal.

Gail and Simon present simulation data indicating, for the data they present in their Table 3, that their method has 62.7% power of rejecting the null hypothesis of no qualitative interaction.

Ciminera et al. used their method to evaluate the data in the examples presented by Gail and Simon and compared their results. The “pushback” method showed substantial evidence of qualitative interaction for the data in both Tables 2 and 3 of Gail and Simon. The method of Gail and Simon had a significance level of <0.05 for Table 3 and <0.10 for Table 2. Ciminera et al. also state (p. 1044) that “In our experience, we have not found a multiclinic study that showed substantial evidence of *qualitative* treatment-by-centre interaction using our procedure. This was especially true of multiclinic studies involving many clinics and comparing equally active (or inactive) treatments.”

3.9 Summary

The methods of detecting qualitative interaction proposed by Ciminera, Heyse, Nguyen and Tukey (1993), Gail and Simon (1985) and Azzalini and Cox (1984) can provide valuable insights into the interpretation of data from multicenter trials. However the operating characteristics of these three methods have not been determined or compared. The utility of the three methods would be enhanced if their characteristics were more completely understood.

CHAPTER IV

EVALUATION OF CURRENT METHODS OF IDENTIFYING OF QUALITATIVE INTERACTION

4.1 Detection of Interaction

Azzalini and Cox (1984), Gail and Simon (1985) and Ciminera, Heyse, Nguyen and Tukey (1993), have presented statistical methods that can be used to identify the presence of qualitative interaction in multicenter trials. Each method also provides a measure of the evidence supporting the existence of a qualitative interaction. The approach, the underlying assumptions and the theory of the three methods differ. Before examining each method individually and comparing the operating characteristics of the three methods, it is important to first review their underlying assumptions and determine a common basis for comparison of the methods.

4.2 Comparison of the Underlying Assumptions of the Methods

4.2.1 Ciminera, Heyse, Nguyen and Tukey

Ciminera and his co-authors do not explicitly discuss any assumptions on the data that are necessary for their test to be valid. However, they do present their method in the context of the analysis of variance model presented in Chapter II, with center as a fixed factor. Consistent with this context, the authors implicitly define the sample estimate of the response of the patients receiving treatment i at center j ($\hat{\mu}_{ij}$) as the mean response, \bar{y}_{ij} . However, they do not use the analysis of variance estimate

of σ^2 , i.e. s_A^2 . Ciminera et al. pool the internally estimated variances of the within center differences between treatments. Their method of variance estimation (p. 1034) “make(s) allowance for the differing estimated variances of the centre means”.

Their method evaluates qualitative interaction between two treatment groups at multiple centers. However, they do not extend their method to the case of more than two treatment groups.

4.2.2 Gail and Simon

Gail and Simon (1985) assume that two treatments are being compared and that the true treatment differences for each subgroup, δ_j , have estimates which are independent and normally distributed with mean δ_j and known variance $\sigma_{d(j)}^2$. Their method is presented in a context more general than that of the analysis of variance. Hence, the δ_j and the estimates of δ_j , d_j , are not restricted to the differences between two means. Gail and Simon state (p. 364) that “the asymptotic normality of most estimators of relative treatment efficacy renders this test broadly applicable.”

The authors note that although the procedures assume known variance $\sigma_{d(j)}^2$, they are valid for large samples if consistent estimates of $\sigma_{d(j)}^2$ are used instead. They also indicate that their methods work for independent d_j with possibly unequal variances. Their method is designed for independent estimates of d_j ; but is also applicable if the d_j are estimated using a model fitted to the entire dataset, if the estimates of the d_j are uncorrelated.

We note that their method is applicable to subgroups of patients other than centers. Their method evaluates qualitative interaction between two treatment groups across multiple subgroups. However, they do not extend their method to the case of more than two treatments.

4.2.3 Azzalini and Cox

The underlying assumptions for the test proposed by Azzalini and Cox (1984) are more extensive, and restrictive, than those of the other methods. Azzalini and Cox assume the normal-theory linear model, with an error variance that is known. A known error variance implies (p. 335) “in effect that the number of degrees of freedom for estimating error is large”. They use the mean square within cells (MSE) as their estimator of the error variance. We have called this estimator s_A^2 . The authors note that the effect of using an estimate of σ^2 has been briefly investigated by simulation. They conclude that, “If the degrees of freedom for estimating σ are at least 30, there is little effect on the significance levels. If the degrees of freedom are between 10 and 30, an apparent 5 per cent level is at worst 10 per cent and an apparent 1 per cent level is at worst 3 per cent.” Hence if σ is estimated from a sample size of less than 30, then the probability of detecting a qualitative interaction could be somewhat higher than the nominal level.

Also implicit in their method is the definition of the sample estimate of the response of the patients receiving treatment i at center j as the cell mean, i.e. $\hat{\mu}_{ij} = \bar{y}_{ij}$.

The test they propose and the distributional calculations for the test statistic are based on the global null hypothesis of no treatment effects. They show that the test is conservative in the presence of main effects.

The authors do not mention any limit on the number of treatment groups that can be simultaneously tested for qualitative interaction with their test. Thus, in contrast to the two previously discussed methods, this method allows simultaneous evaluation of qualitative interaction among multiple treatment groups at multiple

centers. However, the assumption of equal sample sizes across treatments and centers, i.e. $n_{ij} = n$, is an important element of this test.

4.3 Establishing a Common Basis for Comparison of the Three Methods

The major objective of the present research is to compare the operating characteristics of these three methods. In order for such a comparison to be useful, the underlying conditions of the tests must be similar.

We will initially assume that the data used to compare the three methods are consistent with the analysis of variance model presented in Chapter III. The mean $\hat{\mu}_{ij} = \bar{y}_{ij} = \sum_k y_{ijk} / n_{ij}$ will be the measure of treatment response. The analysis of variance model assumed for the Azzalini and Cox and Ciminera et al. methods implies that the effect of the treatment will be evaluated using the mean as the measure of response. The Gail and Simon method does not require this assumption, but the use of this response variable is in harmony with the assumptions of the method. This model also assumes a common variance, σ^2 , estimated by the analysis of variance. We will compare the methods using this estimate of a common variance, which we have termed s_A^2 . This is an additional restriction to the methods of Gail and Simon and Ciminera et al., which allow the variances of the between treatment differences to be heterogeneous across centers.

The methods presented by Ciminera et al. and Gail and Simon are limited to the comparison of interaction between two treatment groups, while the method of Azzalini and Cox can accommodate more than two groups. In the present study, we will only examine comparisons between two treatment groups; for simplicity, they be denoted as treatments 1 and 2, i.e. $i = 1, 2$.

(Note: As previously designated, for simplicity of notation, we will use δ_j and d_j to denote the difference and the estimated difference, respectively, between two designated treatment responses at center j . These quantities could be labeled more precisely as $\delta_{(1, 2; j)}$ and $d_{(1, 2; j)}$. Similarly, we will use a simplistic notation for variances of the estimated treatment differences and the variance estimators)

The issue of uniformity of sample size across treatment groups and centers is very important. The methods presented by Ciminera et al. and Gail and Simon are not restricted by any sample size assumptions across treatment groups or centers. However, the assumption of equal sample sizes across treatment groups and centers is inherent in the derivation of the Azzalini and Cox method. The method as presented requires a common estimate of the variance of the difference in responses of the two treatment groups at center j , which implies a common sample size. This sample size restriction limits the utility of this method in the evaluation of data from multicenter trials. It is common for multicenter trials to have unequal patient numbers across centers, due to differing demographic and recruitment factors. The numbers of patients within a respective center in the various treatment groups may also vary, sometimes as a consequence of the experimental design.

Important assumptions that differ among the three methods and the solutions used to establish a common basis for the simulations are summarized in Table 2.

To broaden the applicability of the results of the present research, we propose an extension of the Azzalini and Cox method to allow unequal sample sizes. We also provide adaptations to the method of Ciminera et al. and the test of Gail and Simon that allow these methods to be used with a common estimate of variance. A discussion of these modifications is presented below.

Table 2

Comparison of Some Important Assumptions of the Three Methods for Identifying Qualitative Interaction

Assumption	Method			Solution for Simulation
	Ciminera et al.	Gail and Simon	Azzalini and Cox	
Measure of Response	Mean	No Assumption	Mean	Use Mean
Number of Treatment Groups Compared	Two Only	Two Only	No Assumption	Use Two Groups Only
Equal Sample Size	No Assumption	No Assumption	Required	Extension of Azzalini and Cox
Common Variance	No Assumption	No Assumption	Required	Adapt Methods to use a Common Variance

4.4 Extension of the Azzalini and Cox Method to the Case of Unequal Sample Sizes Across Centers

The method presented by Azzalini and Cox assumes a common estimate of the variance of the difference in responses of the two treatment groups at all centers, which requires equal sample sizes across treatment groups and centers. They calculate this common variance as

$$(4.1) \quad \sigma_d^2 = 2\sigma^2/n, \text{ where}$$

n is the number of observations within each cell.

Their test is derived from the inequalities that define qualitative interaction,

$$(4.2) \quad d_{j'} \geq t\sigma_d \text{ and } d_{j''} \leq -t\sigma_d,$$

which use the expression of σ_d^2 based on equal sample sizes.

The authors do not present or discuss any results for cases where the sample size may vary across centers and/or treatment groups. We propose an extension of this method to the case of unequal sample sizes. We will examine two approaches for this extension, one is an exact approach and one is an approximate approach.

4.4.1 Exact Approach

We will begin with the exact approach. When the sample size is allowed to vary across centers and treatment groups, the variance of the difference in responses of the two treatment groups may vary for each center j . Under those conditions, for the j th center, (3.39) becomes

$$(4.3) \quad \sigma_{d(j)}^2 = \sigma^2 [(1/n_{1j}) + (1/n_{2j})].$$

It follows that for two centers $j = j', j''$, (4.2) becomes,

$$(4.4) \quad d_{j'} \geq t\sigma_{d(j')} \text{ and } d_{j''} \leq -t\sigma_{d(j'')}.$$

which can be rewritten as,

$$(4.5) \quad d_{j'} \geq t\sigma \sqrt{[(1/n_{1j'}) + (1/n_{2j'})]} \text{ and } d_{j''} \leq -t\sigma \sqrt{[(1/n_{1j''}) + (1/n_{2j'')}]},$$

If we combine the terms that are center-dependent on one side of each inequality, then we have,

$$(4.6) \quad \sqrt{(1/[(1/n_{1j}) + (1/n_{2j})])} d_j \geq t\sigma \text{ and } \sqrt{(1/[(1/n_{1j'}) + (1/n_{2j'})])} d_{j'} \leq -t\sigma.$$

If we substitute the estimator of σ^2 , s_A^2 , into this expression, then (4.6) becomes,

$$(4.7) \quad \sqrt{(1/[(1/n_{1j}) + (1/n_{2j})])} d_j \geq t s_A \text{ and } \sqrt{(1/[(1/n_{1j'}) + (1/n_{2j'})])} d_{j'} \leq -t s_A.$$

The calculation of the test statistic requires finding quadruples of mean responses with pairwise differences having differing signs. That is, all pairs of within center treatment differences, d_j , $d_{j'}$, are examined to find those pairs where the sign of d_j differs from that of $d_{j'}$.

For each pair of differences where the sign of d_j differs from that of $d_{j'}$, the absolute values of the two differences adjusted for sample size (i.e. $\sqrt{(1/[(1/n_{1j}) + (1/n_{2j})])} d_j$ and $\sqrt{(1/[(1/n_{1j'}) + (1/n_{2j'})])} d_{j'}$) are compared and the smaller of the two absolute values, say $d^*_{j,j'}$, is selected,

$$(4.8) \quad d^*_{j,j'} = \min (|\sqrt{(1/[(1/n_{1j}) + (1/n_{2j})])} d_j|, |\sqrt{(1/[(1/n_{1j'}) + (1/n_{2j'})])} d_{j'}|),$$

where the sign of d_j differs from that of $d_{j'}$.

The smaller absolute differences from all such pairs ($d^*_{j,j'}$) are then compared. The largest of all these smaller absolute differences is determined,

$$(4.9) \quad w = \max (d^*_{j,j'}).$$

This maximum difference (w) is used in the calculation of the test statistic. The test statistic is

$$(4.10) \quad t^* = w / (s_A).$$

The level of significance of t^* is

$$(4.11) \quad 1 - \exp [-0.5c(c-1)b(b-1) \{\Phi(-t^*)\}^2], \text{ where}$$

c is the number of centers,

b is the number of treatment levels, and

$\Phi(t^*)$ is the cdf from the standard unit normal distribution of (t^*) .

If the value from (4.11) is less than the predetermined α , then the null hypothesis of no qualitative interaction is rejected and the presence of at least one quadruple with $\mu_{1j'} - \mu_{2j'} > 0$ and $\mu_{1j''} - \mu_{2j''} < 0$ is confirmed.

For the special case where the sample sizes differ across centers, but are equal for the treatment groups within a center j , $n_j = n_{ij} = n_{ij'}$, (4.3) can be simplified as,

$$(4.12) \quad \sigma_{d(j)}^2 = 2\sigma^2 / n_j.$$

Equation (4.7) then becomes

$$(4.13) \quad \sqrt{n_j} d_{j'} \geq t_{s_A} \sqrt{2} \text{ and } \sqrt{n_j} d_{j''} \leq -t_{s_A} \sqrt{2}.$$

For each pair of differences where the sign of $d_{j'}$ differs from that of $d_{j''}$, the absolute values of the two differences adjusted for sample size (i.e. $\sqrt{n_j} d_{j'}$ and $\sqrt{n_j} d_{j''}$) are compared and the smaller of the two absolute values, say $d_{j'j''}^*$, is selected,

$$(4.14) \quad d_{j'j''}^* = \min (|\sqrt{n_j} d_{j'}|, |\sqrt{n_j} d_{j''}|),$$

where the sign of $d_{j'}$ differs from that of $d_{j''}$.

The smaller absolute differences from all such pairs ($d_{j'j''}^*$) are then compared. The largest of all these smaller absolute differences is determined,

$$(4.15) \quad w = \max (d^*_{j,j''}).$$

This maximum difference (w) is used in the calculation of the test statistic.

The test statistic is

$$(4.16) \quad t^* = w / (s_A \sqrt{2}).$$

The level of significance of t^* is

$$(4.17) \quad 1 - \exp [-0.5c(c-1)b(b-1) \{\Phi(-t^*)\}^2], \text{ where}$$

c is the number of centers,

b is the number of treatment levels, and

$\Phi(t^*)$ is the cdf from the standard unit normal distribution of (t^*) .

If the value from (4.17) is less than the predetermined α , then the null hypothesis of no qualitative interaction is rejected and the presence of at least one quadruple with $\mu_{1,j'} - \mu_{2,j'} > 0$ and $\mu_{1,j''} - \mu_{2,j''} < 0$ is confirmed.

The disadvantage of this solution to the problem of unequal sample sizes is that in the calculation of the test statistic, the effect of the size of the difference at a given center is confounded with the influence of the sample sizes at the respective center.

4.4.2 Approximate Approach

A second method of calculating the test statistic in the presence of unequal sample sizes does not have this disadvantage. This method uses an “average” sample size in place of the true sample size, n , in (3.39). The disadvantage of this approach is that it is approximate, whereas the first approach proposed above is an exact approach.

The use of an “average” sample size in the calculation of tests to compare means from an experimental design with different sample sizes has been examined in the context of multiple comparison tests. This current problem is analogous to the problem of determining a common sample size for multiple comparison procedures of means based on unequal sample sizes. This determination has spawned a rich literature and numerous proposed solutions. The interested reader is referred to Hochberg and Tamhane (1987) or Hsu (1996).

One proposal for an “average” sample size is to use the harmonic mean of the sample size for all means in the group of means to be compared. Winer (1971) and Snedecor and Cochran (1967), among others, have suggested the use of this method. The use of this method is not without criticism or caution. Winer (1971) proposes the use of the harmonic mean (p. 216) “if the n_j ’s do not differ markedly from each other”. Although the use of alternative methods may be justified, we will use the harmonic mean in the calculation of a common variance for treatment-by-center cells (\bar{y}_{ij}).

We want to use σ^2 and the harmonic mean to designate a common variance for the difference of means with possibly different sample sizes (n_{ij}). We modify (3.39) to be a common variance of the difference of means,

$$(4.18) \quad \sigma_{d*}^2 = 2\sigma^2 / n^*,$$

based on n^* , the harmonic mean of the sample sizes across all treatments ($i = 1$ to b) and centers ($j = 1$ to c), where

$$(4.19) \quad n^* = bc / (1/n_{11} + 1/n_{12} + \dots + 1/n_{bc}).$$

The calculation of the test follows as described in Chapter III. We use n^* , the estimate of the common variance for the difference of means based on the estimator s_A^2 (from (4.18)) and the following inequality,

$$(4.20) \quad d_{j'} \geq t \sqrt{(2s_A^2 / n^*)} \text{ and } d_{j''} \leq -t \sqrt{(2s_A^2 / n^*)}.$$

For each pair of differences where the sign of $d_{j'}$ differs from that of $d_{j''}$, the absolute values of the two differences are compared and the smaller of the two absolute values, say $d_{j'j''}^*$, is selected,

$$(4.21) \quad d_{j'j''}^* = \min (|d_{j'}|, |d_{j''}|),$$

where the sign of $d_{j'}$ differs from that of $d_{j''}$.

The smaller absolute differences from all such pairs ($d_{j'j''}^*$) are then compared. The largest of all these smaller absolute differences is determined,

$$(4.22) \quad w = \max (d_{j'j''}^*).$$

This maximum difference (w) is used in the calculation of the test statistic.

The test statistic is

$$(4.23) \quad t_A^* = w / \sqrt{(2s_A^2 / n^*)}.$$

Note that t_A^* is an approximation of t^* , based on the approximate estimate of the common variance ($2s_A^2 / n^*$), derived from the harmonic mean.

The level of significance of t_A^* is

$$(4.24) \quad 1 - \exp [-0.5c(c-1)b(b-1) \{\Phi(-t_A^*)\}^2], \text{ where}$$

c is the number of centers,

b is the number of treatment levels, and

$\Phi(t_A^*)$ is the cdf from the standard unit normal distribution of (t_A^*).

If the value from (3.48) is less than the predetermined α , then the null hypothesis of no qualitative interaction is rejected and the presence of at least one quadruple with $\mu_{1j^*} - \mu_{2j^*} > 0$ and $\mu_{1j^{**}} - \mu_{2j^{**}} < 0$ is confirmed.

In the case of equal sample sizes across treatments and centers, the exact and approximate methods are equivalent and both are equivalent to the equal sample size method given by Azzalini and Cox, as described in Chapter III.

4.5 Adaptation of the Methods of Ciminera et al. and Gail and Simon to Incorporate the Variance Estimate From Analysis of Variance

The recommended procedures for the method of Ciminera et al. use the pooled overall estimate across centers of the internally estimated between treatment variances. However, the use of the estimate of the common variance estimated by the MSE from analysis of variance is not inconsistent with the use of a pooled variance estimator proposed by the authors.

The pooled overall estimate across centers of the between treatment variance used by Ciminera et al. is calculated as described below (4.25)

$$(4.25) \quad s_{P \text{ overall}}^2 = [(\sum_j v_{d(j)} s_{P d(j)}^2) / \sum_j v_{d(j)}], \text{ where}$$

$s_{P d(j)}^2$ is the internally estimated variance of the difference in the mean responses of the two treatment groups at center j , and
 $v_{d(j)} = n_{1j} + n_{2j} - 2$.

We can replace the estimator $s_{P d(j)}^2$ in this equation with the estimate of the variance of the difference in the mean responses of the two treatment groups at center j derived from the analysis of variance estimate of the common variance,

$$(4.26) \quad s_{A d(j)}^2 = s_A^2 [(1 / n_{1j}) + (1 / n_{2j})].$$

In the special case where the sample sizes differ across centers, but are equal for the treatment groups within a center j , $n_j = n_{1j} = n_{2j}$, (3.12) can be simplified as,

$$(4.27) \quad s_{A\ d(j)}^2 = 2s_A^2 / n_j.$$

With the substitution of $s_{A\ d(j)}^2$ from (3.12) or (4.27), (4.25) becomes

$$(4.28) \quad s_{P\ overall}^2 = [\sum_j v_{d(j)} s_{A\ d(j)}^2 / \sum_j v_{d(j)}], \text{ where}$$

$s_{A\ d(j)}^2$ is the estimate of the variance of the difference in the mean responses of the two treatment groups at center j based on the common estimate of the variance from analysis of variance, and
 $v_{d(j)} = n_{1j} + n_{2j} - 2$.

Using this estimator of the pooled overall estimate across centers of the between treatment variance, the calculations for the method of Ciminera et al can proceed as described in Chapter III.

Gail and Simon are not explicit in their description of the variance estimator to use in the calculation of their method. They do provide the following specifications.

They assume that two treatments are being compared and that the true treatment differences for each subgroup, δ_j , have estimates which are independent and normally distributed with mean δ_j and known variance $\sigma_{d(j)}^2$. They state that consistent estimates of $\sigma_{d(j)}^2$ may be inserted in all of the equations presented above without altering the asymptotic distribution theory and that the procedures they introduce are valid for large samples if consistent estimates of $\sigma_{d(j)}^2$ are used. They also indicate that their method works for independent d_j with possibly unequal variances. Their method is designed for independent estimates of d_j ; but is also applicable if the d_j are estimated using a model fitted to the entire dataset, if the estimates of the d_j are uncorrelated. Gail and Simon also state (p. 364) that “the

asymptotic normality of most estimators of relative treatment efficacy renders this test broadly applicable.”

The use of the estimate of the common variance estimated by the MSE from analysis of variance is consistent with the stated assumptions of the authors.

Recall that for the methods of Gail and Simon, the quantities Q^- and Q^+ are calculated as follows:

$$(4.29) \quad Q^- = \sum_j (d_j^2 / \sigma_{d(j)}^2) I(d_j > 0)$$

and

$$(4.30) \quad Q^+ = \sum_j (d_j^2 / \sigma_{d(j)}^2) I(d_j < 0), \text{ where}$$

$I(d_j > 0) = 1$ if $d_j > 0$ and 0 otherwise, and
 $I(d_j < 0) = 1$ if $d_j < 0$ and 0 otherwise.

If we use the estimate of the common variance derived from the analysis of variance, and substitute $s_{A \text{ dif}}^2$ from (3.12) or (4.27), then (3.30) and (3.31) become, respectively,

$$(4.31) \quad Q^- = \sum_j (d_j^2 / s_{A \text{ dif}}^2) I(d_j > 0)$$

and

$$(4.32) \quad Q^+ = \sum_j (d_j^2 / s_{A \text{ dif}}^2) I(d_j < 0).$$

The hypothesis of qualitative interaction and the hypothesis that one treatment is at least as good as the other treatment can be tested using these estimates of Q^- and Q^+ , as described in Chapter III.

The estimator $s_{A \text{ dif}}^2$ can also be used in the test of no interaction, based on the weighted residual sum of squares,

$$(4.33) \quad H = \sum_j (d_j - \bar{d})^2 / s_A^2 d(j), \text{ where}$$

$$(4.34) \quad \bar{d} = [\sum_j (d_j / s_A^2 d(j))] / [\sum_j (1 / s_A^2 d(j))] \text{ is a weighted mean.}$$

Under the hypothesis of no quantitative interaction, H has a central chi-square distribution with degrees of freedom, $df = c - 1$, where c is the number of subgroups.

4.6 Examination of the Signs of the Treatment Effects in the Raw Data

In addition to comparing the three methods to each other, we will also compare them to an ad-hoc method for detection of qualitative interaction. A common approach used to identify qualitative interaction is to visually examine the means and the treatment differences across centers, either in a tabular or graphical presentation. The basis of the method is that the presence of treatment effects with differing directions across centers is positive evidence of qualitative interaction. Hence, if the signs of the treatment effects are not all positive or all negative, then the study is considered to show “substantial evidence” of qualitative interaction, based on this criterion. This method is similar to the method of Ciminera et al., except that the signs of the treatment effects from the raw data are examined, rather than the “pushed back” treatment effects.

4.7 Example

The four methods, including the method of examining the signs in the raw data, will be illustrated with an example that incorporates the use of the common estimate of the variance and has unequal sample sizes requiring the extensions of the Azzalini and Cox method described above. Both the exact and approximate extensions of the Azzalini and Cox method will be illustrated.

4.7.1 Introduction and Results of the Analysis of Variance

A multicenter trial with two treatment arms was conducted at three centers. The numbers of patients enrolled varied across centers, but was consistent for both treatments within a center. Summary statistics for patient response for each treatment group and for the difference between treatment groups at each center are presented in Table 3.

The data were analyzed using the SAS MIXED procedure with the following statements (plus additional statements to create the output datasets needed to compute the results for the interaction methods):

```
proc mixed data = &dataset;
  class center treat;
  model &variable = center treat center*trt;
  .
  .
run;
```

The estimate of the residual was 0.6408, with 60 degrees of freedom. The results of the tests of the fixed effects of the model are presented in Table 4.

The F value for the test of the treatment effect was 4.95, with one degree of freedom, which has a significance level of 0.0299. The F value for the test of the treatment-by-center interaction was 7.00, with two degrees of freedom, which has a significance level of 0.0018.

Hence from the analysis of variance we conclude that there is a significant difference between the mean response for the two treatment groups ($p = 0.0299$). However, the presence of a significant ($p < 0.0018$) treatment-by-center interaction must be further investigated before the treatment difference can be confidently interpreted.

Table 3
Summary Statistics for Patient Response

Center	Statistic	Treatment 1	Treatment 2	Difference
1	n	8	8	8
	Mean	5.095	4.64	0.455
2	n	10	10	10
	Mean	6.172	6.618	-0.446
3	n	15	15	15
	Mean	6.713	8.083	-1.369

Table 4
Tests of Fixed Effects

Source	Degrees of Freedom	Type III F Statistic	Pr > F
Site	2	52.30	0.0001
Treatment	1	4.95	0.0299
Site X Treatment	2	7.00	0.0018

The analysis of variance estimate of σ^2 , s_A^2 , is 0.6408. The unequal sample sizes across the centers prohibit the calculation of a common $\sigma_{d(j)}^2$. However, we can calculate an estimate of $\sigma_{d(j)}^2$ for each center. From the variance estimate, s_A^2 , we can calculate $s_{A\ d(j)}^2$ for each of three centers. Since, for each center, $n_j = n_{1j} = n_{2j}$, we can use (4.12),

$$(4.35) \quad s_{A d(1)}^2 = 2 (0.6408 / 8) = 0.1602$$

$$s_{A d(2)}^2 = 2 (0.6408 / 10) = 0.1282$$

$$s_{A d(3)}^2 = 2 (0.6408 / 15) = 0.0854.$$

4.7.2 Signs of the Treatment Effects in the Raw Data

From Table 3, we note that the mean treatment effects (differences) for Centers 2 and 3 are negative, whereas the sign of the effect at Center 1 is positive. Since the signs of the three treatment effects do not all have the same sign, we conclude that, based on this procedure, that there is “substantial” evidence for the existence of qualitative interaction.

4.7.3 Azzalini and Cox Exact Method

We will next test for qualitative interaction using the method of Azzalini and Cox. Since the sample sizes are not equal across all treatment groups and centers, we will use the extensions of the method presented above. We will first use the exact method.

The first step in the calculation of the Azzalini and Cox test for qualitative interaction is the examination of all pairs (i.e. across centers) of treatment differences to determine if the differences have the same signs. In this example there are three pairs to examine: center 1 and center 2; center 1 and center 3 and center 2 and center 3. We note (in Table 3) that 1 and 2 have differing signs as do 1 and 3, while 2 and 3 are both negative. For each pair with differing signs we calculate the minimum absolute value, adjusted for sample size (4.14),

$$(4.36) \quad d^*_{1,2} = \min (|\sqrt{(8)}*0.455|, |\sqrt{(10)}*-0.446|) = 1.287,$$

$$(4.37) \quad d_{1,3}^* = \min (|\sqrt{(8)*0.455}|, |\sqrt{(15)*-1.369}|) = 1.287.$$

The calculation of w requires us to take the maximum of these minimums
(4.15),

$$(4.38) \quad w = \max (1.287, 1.287) = 1.287.$$

We divide this maximum by $s_A\sqrt{2}$ (4.16) to get our test statistic t^* ,

$$(4.39) \quad t^* = 1.287 / (0.8005*1.414) \\ = 1.137.$$

The level of significance of t^* is

$$(4.40) \quad 1 - \exp [-0.5*3(3-1)2(2-1) \{\Phi(-1.137)\}^2] = 0.093$$

We conclude that, based on this test, the probability that the qualitative interaction present in this study is due to chance alone is 0.093.

4.7.4 Azzalini and Cox Approximate Method

We will now calculate the test using the approximate method of Azzalini and Cox. The calculation of a common estimate of the variance of the means requires the use of the harmonic mean, n^* (4.19),

$$(4.41) \quad n^* = 2*3 / (1/ 8 + 1/ 8 + 1/ 10 + 1/ 10 + 1/ 15 + 1/ 15) \\ = 10.29.$$

For each pair with differing signs we calculate the minimum absolute value
(4.21),

$$(4.42) \quad d^*_{1,2} = \min (|0.455|, |-0.446|) = 0.446,$$

$$(4.43) \quad d^*_{1,3} = \min (|0.455|, |-1.369|) = 0.455.$$

The calculation of w requires us to take the maximum of these minimums

(4.22),

$$(4.44) \quad w = \max (0.446, 0.455) = 0.455.$$

We divide this maximum by the square root of the common estimate of the variance of the means to get our test statistic t_A^* (4.23),

$$(4.45) \quad t_A^* = 0.455 / \sqrt{(2 \cdot 0.6408 / 10.29)} \\ = 1.289.$$

The level of significance of t_A^* is

$$(4.46) \quad 1 - \exp [-0.5 \cdot 3(3-1)2(2-1) \{\Phi(-1.289)\}^2] = 0.0568.$$

We conclude that, based on this test, the probability that the qualitative interaction present in this study is due to chance alone is 0.0568.

4.7.5 Gail and Simon

The calculation of the Gail and Simon procedures begins with determination of which within center treatment differences have positive and negative signs. We note that the sign of difference from center 1, d_1 , is positive, while the signs of the differences from centers 2 and 3, d_2 and d_3 , have negative signs. We first calculate Q^+ (4.32), which sums over all $d_j < 0$ (i.e. centers 2 and 3),

$$(4.47) \quad Q^+ = [(-0.446)^2 / (0.1282)] + [(-1.369)^2 / (0.0854)] \\ = 23.487.$$

Likewise, we calculate Q^- (4.31), which sums over all $d_j > 0$ (i.e. center 1),

$$(4.48) \quad Q^- = (0.455)^2 / (0.1602) \\ = 1.2923.$$

The test statistic, Q , is the minimum of Q^- and Q^+ ,

$$(4.49) \quad Q = \min (1.2923, 23.487) \\ = 1.2923.$$

The level of the test for this critical value is 0.292, based on (3.33). We conclude that, based on this test, the qualitative interaction present in this study is not significant at a test level of 10%.

Using the statistic Q^- we can also test the hypothesis that the mean response for patients in Treatment 2 is at least as good as the mean response for patients in Treatment 1. The level of the test for this critical value is 0.3838 (from 3.35), which is greater than 0.025. Thus, we cannot reject the null hypothesis that the mean response for patients in Treatment 2 is at least as good as the mean response for patients in Treatment 1.

Likewise, using the statistic Q^+ we can also test the hypothesis that the mean response for patients in Treatment 1 is at least as good as the mean response for patients in Treatment 2. The level of the test for this critical value is < 0.0001 (from 3.35), which is less than 0.025. We can reject the null hypothesis that the mean response for patients in Treatment 1 is at least as good as the mean response for patients in Treatment 2.

From the results of these tests we can conclude, at a level of 0.025, that the mean response for Treatment 2 is equal to or greater than the mean response for Treatment 1 at every site. On the other hand, the mean response for patients in Treatment 1 is not equal to or greater than the mean response for patients in Treatment 2 at all sites.

The test for interaction proposed by Gail and Simon requires the calculation of a weighted mean, \bar{d} (4.34),

$$\begin{aligned} (4.50) \quad \bar{d} &= [(0.455 / 0.1602) + (-0.446 / 0.1282) + (-1.369 / 0.0854)] / \\ &\quad [(1 / 0.1602) + (1 / 0.1282) + (1 / 0.0854)] \\ &= -0.6471. \end{aligned}$$

A statistic, H , is then calculated,

$$\begin{aligned} (4.51) \quad H &= [(0.455 - (-0.6471))^2 / 0.1602] + [(-0.446 - (-0.6471))^2 / 0.1282] + \\ &\quad [(-1.369 - (-0.6471))^2 / 0.0854] \\ &= 13.9970. \end{aligned}$$

Under the hypothesis of no quantitative interaction, H has a central chi-square distribution with degrees of freedom, $df = 2$; for a value of 13.9970, $p < 0.005$.

From these three tests of Gail and Simon, we can conclude that there is significant treatment-by-center interaction ($p < 0.005$), that the interaction is not significant qualitative interaction ($p = 0.29$), and that the mean response for Treatment 2 is at least as good as the mean response for Treatment 1 ($p < 0.0001$).

4.7.6 Ciminera et al.

The “pushback” method of Ciminera et al. requires the calculation of a pooled overall estimate across centers of the between treatment variance (4.28),

$$\begin{aligned}
 (4.52) \quad s_P^2_{\text{overall}} &= [(14 * 0.1602) + (18 * 0.1282) + (28 * 0.0854)] / \\
 &\quad [14 + 18 + 28] \\
 &= 0.1157.
 \end{aligned}$$

The weighted median difference across centers (m) is calculated by taking the median of the vector of mean treatment differences, constructed by including the mean treatment difference at each center once for each patient in the trial at the center. In this case, we have (starting with the smallest value) 30 observations with the value of -1.369 , 20 observations with a value of -0.446 and 16 observations with a value of 0.455 . The median of this weighted vector of mean differences is the average of the 33rd and 34th ordered observations of the vector,

$$\begin{aligned}
 (4.53) \quad m &= [-0.446 + (-0.446)] / 2 \\
 &= -0.446.
 \end{aligned}$$

The median difference (m) is subtracted from the mean treatment difference from each center (d_j) to determine a deviation for each center. Each deviation is then standardized by dividing by the pooled standard deviation from (4.52). The calculation for center 1 is shown below, values for all three centers are displayed in Table 5,

$$\begin{aligned}
 (4.54) \quad D_{s(1)} &= [0.455 - (-0.446)] / \sqrt{0.1157} \\
 &= 2.6489.
 \end{aligned}$$

When the d_j of the centers are ordered and ranked, we have the ranks presented in Table 5. For centers 2 and 3, with d_i equal to or below the overall median difference (-0.446), the lower 10 per cent point of the incomplete beta function is calculated using parameters v_1 and v_2 shown in Table 5. For example for center 3,

$$(4.55) \quad P_3 = 10^{\text{th}} \text{ percentile of } \beta(1, 3) \\ = 0.0345.$$

For center 1, which had a d_j above the median (-0.446), the upper 10 per cent point (i.e. the 90 per cent point) of the incomplete beta function is calculated using parameters: v_1 and v_2 ,

$$(4.56) \quad P_1 = 90^{\text{th}} \text{ percentile of } \beta(3, 1) \\ = 0.9655.$$

The percent points of the incomplete beta function calculated for each center are then used as percentiles of the Student t distribution and the t value (t_j) corresponding to each value is determined. For each center, t_j is the value that gives a cumulative cdf for the t-distribution, with the respective df, equal to the corresponding P_j previously calculated. All t_j are presented in Table 5, the calculation of t_1 is shown below:

$$(4.57) \quad t_1 = t(0.9655, 2(8 - 1)).$$

Therefore,

$$(4.58) \quad t_1 = 1.9694.$$

The t_j value calculated for each center is then subtracted from the standardized deviation for the center. For example, for center 1,

$$(4.59) \quad D_{p(1)} = 2.6489 - 1.9694 \\ = 0.6795.$$

However, since zero is always pushed back to zero, the value for center 2, $D_{p(2)}$, is 0.0.

The pushed-back standardized deviations are restored to their original scale (destandardized) by multiplying by the pooled standard deviation and adding the overall median difference (m). For example, for center 1,

$$(4.60) \quad D_{d(1)} = (0.3401 * 0.6795) + (-0.446) \\ = -0.2149.$$

From Table 5, we note that the destandardized values ($D_{d(j)}$) for all three centers are negative. Since the signs of the three “pushed back”, destandardized values all have the same sign, we conclude that, based on this procedure, that there is not “substantial” evidence for the existence of qualitative interaction.

4.7.7 Summary

In summary, there is significant evidence that there is a difference between the two treatment groups ($p = 0.0299$, from ANOVA) and that the mean response for Treatment 2 is at least as good as the mean response for Treatment 1 ($p < 0.0001$, Gail and Simon). There is a significant treatment-by-center interaction ($p = 0.0018$, ANOVA; $p < 0.005$, Gail and Simon H statistic). However, even though the mean treatment differences for the three centers show evidence of qualitative interaction

based on the signs of the treatment effects in the raw data, the three methods of identifying qualitative interaction indicate that the qualitative interaction may not be significant or substantial. If tested at a level of 0.10, the test of Azzalini and Cox provides significant evidence for the existence of qualitative interaction ($p = 0.0933$, exact test; $p = 0.0568$, approximate test). However, the significance level of Gail and Simon's test is > 0.10 and the procedure of Ciminera et al. does not provide substantial evidence for the existence of qualitative interaction. In conclusion, we could justifiably state that Treatment 2 is significantly better than Treatment 1 and that the assertion holds across all centers.

Table 5
Results of the Ciminera et al. Procedure

Center	Between-treatment Difference (d_j)	Rank of d_j	Pooled Variance (s_p^2 overall)	Degrees of Freedom ($v_{d(j)}$)	Standardized Difference ($D_{s(j)}$)	β Function Parameter 1 (v_1)	β Function Parameter 2 (v_2)	P_j	t_j	Pushed-back Value ($D_{p(j)}$)	Destandardized Value ($D_{d(j)}$)
3	-1.369	1	0.1157	28	-2.7135	1	3	0.0345	-1.8909	-0.8226	-0.7258
2	-0.446	2	0.1157	18	0.0000	2	2	0.8042	0.8778	0.00000	-0.446
1	0.455	3	0.1157	14	2.6489	3	1	0.9655	1.9694	0.6795	-0.2149

CHAPTER V

DESCRIPTION OF SIMULATION PROCEDURES FOR COMPARING METHODS OF IDENTIFYING QUALITATIVE INTERACTION

5.1 Introduction

The primary objective of the present study is to determine and compare the operating characteristics of the three methods proposed for the detection of qualitative interaction, i.e. Azzalini and Cox (1984), Gail and Simon (1985) and Ciminera, Heyse, Nguyen and Tukey (1993). The utility of the methods in evaluating qualitative interaction in data from multicenter trials may be affected by many factors (some controllable and some uncontrollable) such as the number of treatment groups, the number of centers, the variance estimator chosen, homogeneity of variance, total sample size, sample size differences among treatment groups and/or centers, the magnitude of the overall treatment difference across all centers, and the patterns of the interaction. Clearly, the examination of all of the factors that can affect the performance of these tests is beyond the scope of this research; however, this study is a step in the evaluation process.

The methodology chosen for the evaluation and comparison of the methods was the analysis of simulated datasets prepared with characteristics that allowed the evaluation and comparison of the methods under specified conditions. The characteristics were chosen to provide insight into the effect of some of the factors described above on the performance of the methods. The selected characteristics will be explained in more detail later in this chapter.

For each specified data pattern (see description of data patterns later in this chapter), 2500 datasets were simulated and analyzed. Each simulated dataset was evaluated using standard analysis of variance methods and the procedures proposed for the detection of qualitative interaction. The presence of treatment-by-center interaction was determined using the analysis of variance test for interaction and using the H statistic of Gail and Simon. Qualitative interaction was evaluated using the methods of Azzalini and Cox, Gail and Simon and Ciminera et al. The “non-inferiority” of one treatment vs. the other was also examined using the test proposed by Gail and Simon.

The power of the methods to detect qualitative interaction was evaluated using datasets with simulation specifications defining the presence of qualitative interaction. That is, the differences of the treatment means specified for these simulation patterns differed in magnitude and in direction between centers. The error rates of the methods were evaluated using datasets that had simulation specifications that defined the absence of qualitative interaction. Datasets without qualitative interaction could have quantitative interaction or could have no interaction.

5.2 Specification of Simulated Datasets

As presented above, there are many factors that can influence the detection of qualitative interaction. A limited number of factors and factor levels were chosen for evaluation in the present study.

The number of treatment groups assessed in a multicenter trial may vary from one to several. In the present study, simulated datasets contain two treatment groups. As previously indicated, the methods of Gail and Simon and Ciminera et al. were developed for the comparison of only two treatment groups. In many cases the

treatment group comparisons in multicenter trials that are ultimately of interest are pairwise comparisons. Thus, the evaluation of qualitative interaction among pairs of treatment groups may be important in trials with more than two treatment groups.

We began our research with the simplest forms of qualitative interaction; we simulated trials with two treatments and two centers. An understanding of the performance of the tests in a two-center design will provide insight into performance in more complex designs. The interaction patterns found in trials with more than two centers may be described by combinations of the two-center interaction patterns. The results of this evaluation will also be applicable to designs where a two-level factor other than center is included in the model. An example would be an experiment with treatment assignments randomized within sex, rather than across centers. In this study, we also evaluate a design with three centers.

As indicated in Chapter IV, we examined data generated under the assumptions of a normal-theory linear model, with the e_{ijk} assumed to be independent, $N(0, \sigma^2)$. Without loss of generality, σ^2 was defined to be 1.0 for all datasets. Although the sample variance was assumed to be equal for all treatment groups and centers, the standard errors of the means for the treatment groups could vary across centers as we allowed the sample size to vary across centers. The numbers of patients per treatment group were allowed to vary between centers, but were equal between the two treatment groups at a given center. Varying the sample sizes among centers provided an evaluation of the methods in the presence of unequal sample sizes, as well as equal sample sizes.

Simulation patterns were designed to assess both power and error rates of the examined methods. Patterns were designed to have no interaction, quantitative

interaction or qualitative interaction. Of these patterns, some were designed with treatment effect and some with no treatment effect.

The error rates of the tests for overall interaction were determined using datasets defined to have no interaction. The error rates for the methods of detecting qualitative interaction were determined using datasets defined to have no qualitative interaction. The power of the tests to detect interaction was examined using datasets defined to have interaction, either qualitative or quantitative. The power of the tests to detect qualitative interaction was examined using datasets defined to have qualitative interaction.

The performance of the methods in the presence of a treatment effect was evaluated using data simulated to have a treatment effect. In datasets with a treatment effect, the overall sample size per treatment and the magnitude of the overall treatment effect were specified so that the treatment effect was significant in 80% of the simulated datasets, when tested at a level of 5%. These levels of significance and power correspond to those used to design multicenter trials. The sample size of multicenter trials is frequently chosen to provide 80%, or sometimes 90%, power to detect a difference in the primary efficacy variable, when tested at a two-sided level of 5%.

5.3 Evaluation of Simulated Datasets

5.3.1 Simulation Procedures

For each simulation pattern described below, 2500 datasets were simulated using SAS data statements. The data were generated from a standard normal distribution adjusted to provide treatment means based on the specifications of the respective simulation pattern. The same initial seed number was used for the

simulation of the random numbers in the 2500 datasets that formed the basis for all patterns and sample size configurations. The size of the simulation, 2500 datasets, was determined based on the desired precision of the results of the simulation and a hardware limitation of the equipment used for the simulations.

The SAS code used to generate the data for two centers was:

```
%macro simulate(&seed, &nsim, &s1n, &s2n,
                &s1t1, &s1t2, &s2t1, &s2t2);
data multi (keep = site trt npat x);
  seed2 = &seed + &nsim;
  do site = 1 to 2;
    if site = 1 then enrolled = &s1n;
    else if site = 2 then enrolled = &s2n;
    do trt = 1 to 2;
      do npat = 1 to enrolled;
        call rannor(seed2, x);
        if site = 1 and trt = 1 then x = x + &s1t1;
        if site = 1 and trt = 2 then x = x + &s1t2;
        if site = 2 and trt = 1 then x = x + &s2t1;
        if site = 2 and trt = 2 then x = x + &s2t2;
        output;
      end;
    end;
  end;
run;
%mend simulate;
```

The definitions of the macro variables in the simulation macro are:

&seed = the initial seed number (Note: the same, randomly selected number, 878945, was used for all simulations.),
 &nsim = the index of the simulation, from 1 to 2500,
 &s1n = the sample size for each treatment group at center 1,
 &s2n = the sample size for each treatment group at center 2,
 &s1t1 = the desired mean of the numbers generated for center 1 and treatment 1,
 &s1t2 = the desired mean of the numbers generated for center 1 and treatment 2,
 &s2t1 = the desired mean of the numbers generated for center 2 and treatment 1,
 &s2t2 = the desired mean of the numbers generated for center 2 and treatment 2.

Note that each ϵ_{ijt} is added to a randomly generated number with expected value of zero.

The results of the test of significance of the overall treatment effect were evaluated for each of the 2500 simulated datasets generated for each simulation pattern. The overall treatment effect was considered to be significant if the F statistic for the test had a significance level less than or equal to 5%. The number of significant tests from the 2500 simulated datasets was summarized. For datasets simulated based on a significant treatment effect, the number of significant tests is considered to be the power of the test. For datasets with a non-significant treatment effect pattern, the percentage of significant tests is considered to be the error rate.

5.3.2 Tests of Overall Interaction

The results of the test of significance of the treatment-by-center interaction from analysis of variance were obtained for each simulated dataset and the interaction was considered to be significant if the F statistic for the test had a significance level less than or equal to 10%. The numbers of significant tests from the 2500 simulated datasets are summarized.

The overall treatment-by-center interaction was also tested using the H statistic proposed by Gail and Simon. The test was conducted using SAS data steps and procedures. The interaction was considered significant if the level of the associated chi-square test was less than or equal to 10%. The numbers of significant tests from the 2500 simulated datasets are summarized.

The power and errors rates of the two tests of overall treatment-by-center interaction were compared with each other and with the results of the tests of qualitative interaction presented below. For datasets simulated based on the presence of a treatment-by-center interaction, the number of significant tests is considered to be

the power of the test. For datasets without a treatment-by-center interaction, the number of significant tests is considered to be the error rate.

5.3.3 Methods for the Detection of Qualitative Interaction

Qualitative interaction was evaluated using the tests of Azzalini and Cox and Gail and Simon, and the method of Ciminera et al. Most of the simulated datasets (as described below) had unequal sample sizes, hence the extended methods of Azzalini and Cox presented in Chapter IV were used. Both the exact and approximate methods were evaluated. (In cases of equal sample sizes, both of these methods should produce results equal to the original method. However, for simplicity of programming, the extended methods were used in all cases.) The methods of Gail and Simon and Ciminera et al. were evaluated using the variance estimate from the analysis of variance, as presented in Chapter IV. Results for all methods were calculated using SAS data statements and procedures.

The significance of the Azzalini and Cox test was calculated for each simulated dataset for both the exact and approximate procedures. A dataset was considered to have a significant qualitative interaction by this criterion if the level of the test was less than or equal to 10%. The number of significant tests from the 2500 simulated datasets was summarized.

Each dataset was tested for qualitative interaction using Gail and Simon's test with the minimum of Q^+ and Q^- . The qualitative interaction was considered significant by this criterion if the level of the test was less than or equal to 10%. The number of significant tests from the 2500 simulated datasets was summarized.

The results of the method of Ciminera et al. were examined for each dataset. Any dataset with "pushed back" (destandardized) mean differences which were not all

either positive or negative was considered to show “substantial evidence” of the presence of qualitative interaction. The number of datasets showing “substantial evidence” was summarized.

The signs of mean treatment effects (differences) for each center were also examined in each dataset. If the signs of the treatment differences were not the same for all centers, then the dataset was considered to show “substantial evidence” of the presence of qualitative interaction based on the “raw data”. The number of datasets showing “substantial evidence” was summarized.

The power and error rates of these methods, including both the exact and approximate Azzalini and Cox tests and the raw data method, were compared. Significant and non-significant results of the Azzalini and Cox and Gail and Simon tests were tabulated, as well as the number of tests that showed substantial evidence of qualitative interaction using the Ciminera et al. procedure and the raw data method. For datasets simulated based on the presence of a qualitative treatment-by-center interaction, the number of significant tests (or those showing “substantial evidence”) is considered to be the power of the test for each respective method. For datasets without a qualitative treatment-by-center interaction, the number of significant tests (or those showing “substantial evidence”) is considered to be the error rate for the method.

5.3.4 Test of Non-inferiority

The null hypotheses that response of patients in each treatment group is as least as good as the response in the other treatment group were tested using the test proposed by Gail and Simon. The test for each treatment was considered to be

significant if the level of the test was less than or equal to 2.5%. The number of significant test results for each treatment group was summarized.

5.3.5 Two-stage Test Results

An ultimate objective of the test for overall interaction may be model simplification if no interaction is present. That is, if there is no significant interaction, the interaction term may be removed from the analysis of variance model for the final analysis.

As presented in Chapter II, several authors believe that some level of quantitative interaction is inevitable and it is only the presence of qualitative interaction that clouds the interpretation of results from multicenter trials. Hence the objective of identifying qualitative interaction may be model simplification if no qualitative interaction is present. That is, if there is no significant qualitative interaction or a lack of “substantial” evidence of qualitative interaction, the interaction term may be removed from the analysis of variance model for the final analysis.

The effectiveness of using each of the methods of detecting interaction or qualitative interaction as a pretest for the removal of the interaction term from the model was evaluated. For each dataset with no significant interaction or with no significant or substantial evidence of qualitative interaction, the significance of the overall treatment effect was evaluated using a Type II model (without the interaction term). For each dataset with significant interaction or with significant or substantial evidence of qualitative interaction, the significance of the overall treatment effect was evaluated using a Type III model (which includes the interaction term).

The effectiveness of each method as a pretest was evaluated by determining the rejection rates of the test of significance of the overall treatment effect with a final model analysis model based on using each method as a pretest. These rejection rates were each compared to the rejection rate when the full (Type III) model was always used and to the rejection rate when the reduced (Type II) model was always used. The Type III and Type II treatment means were also calculated as an aid to the interpretation of the models.

5.4 The General Linear Mixed Model

The analysis of variance procedure used in the present study was the general linear mixed model implemented with the MIXED procedure of SAS, version 6.12 (SAS Institute Inc, 1997). The common, overall variance estimate and the analysis of variance tests for the treatment effect and the treatment-by-center interaction were obtained from this procedure. The MIXED procedure was used for analysis of variance for this study since parameter estimates can be easily saved in output datasets for subsequent analyses. Output of parameter estimates into datasets is more difficult when the alternative, GLM procedure is used. (The output datasets are now available in newer versions of SAS; however, that was not the case when the present study was undertaken.) The maximum likelihood estimate of σ^2 was estimated for this study under the assumption that the covariance matrix \mathbf{R} of the general linear mixed model was $\sigma^2\mathbf{I}$. This estimate is analogous to the MSE from a GLM analysis of variance. Although a detailed description of this mixed model approach will not be presented here, some basic details will be discussed.

In matrix notation, the general linear mixed model has the equation,

$$(5.1) \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}, \text{ where}$$

\mathbf{Y} is the vector of observations,
 $\mathbf{X}\boldsymbol{\beta}$ represents the fixed portion of the model,
 $\mathbf{Z}\boldsymbol{\gamma}$ represents the random portion of the model and \mathbf{e} is a vector of errors.

In the case of a fixed effects model, as is the case in the present study, there are no random effects, hence $\mathbf{Z} = \mathbf{0}$, and the model simplifies to,

$$(5.2) \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

The distribution of \mathbf{e} in the general linear mixed model is assumed to be MVN $(\mathbf{0}, \mathbf{R})$. As previously stated, for this study, the e_{ijk} are assumed to be independent, $N(0, \sigma^2)$. Thus $\mathbf{R} = \sigma^2 \mathbf{I}$, where \mathbf{I} is an identity matrix with size equal to the total number of observations. The first step in parameter estimation and calculation of inferential statistics using the SAS MIXED procedure is to estimate the \mathbf{R} matrix using maximum likelihood methods. This procedure produces an estimate of σ^2 , which corresponds to the MSE in a GLM analysis of variance. The parameters of $\boldsymbol{\beta}$ are then estimated using mixed model equations, analogous to the normal equations used to estimate $\boldsymbol{\beta}$ when a GLM approach is used. The hypotheses concerning the treatment effect and the treatment-by-center interaction are tested using hypothesis matrices, \mathbf{L} . Each hypothesis to be tested, for example $H_0: B_1 - B_2 = 0$, can be formulated in terms of ,

$$(5.3) \quad H_0: \mathbf{L}\boldsymbol{\beta} = \mathbf{0}.$$

The hypothesis is tested with an F-statistic calculated as follows,

$$(5.4) \quad F = \hat{\boldsymbol{\beta}}' \mathbf{L}' (\mathbf{L}' (\mathbf{X}' \hat{\mathbf{R}}^{-1} \mathbf{X}) \mathbf{L})^{-1} \mathbf{L} \hat{\boldsymbol{\beta}} / \text{rank}(\mathbf{L}), \text{ where}$$

$\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$, and
 $\hat{\mathbf{R}}$ is the estimate of \mathbf{R} .

This F-statistic has degrees of freedom equal to rank (L), and the degrees of freedom associated with the estimate of σ^2 .

5.5 Simulation of Trials With Two Centers

Datasets with two treatment groups at two centers were simulated. Eleven patterns of datasets were simulated: nine patterns with a difference between treatments (overall treatment effect) and two patterns without a difference between treatments (no overall treatment effect). Patterns were selected to display either no interaction, quantitative interaction or qualitative interaction.

5.5.1 Patterns With a Difference Between Treatments (Overall Treatment Effect)

For these patterns, the overall variance, the overall difference between the means of Treatment 1 and Treatment 2 (the overall treatment effect) across the two centers and the total number of patients per treatment were determined so that the simulated data would have an overall mean treatment effect that would be significant at the 5% level in 80% of simulated datasets (i.e. power of 80%). The overall variance was 1.0, the difference between the means of Treatment 1 and Treatment 2 was 0.5 for the case with equal sample sizes and the sample size was 64 patients per treatment. In other words, with sixty-four patients per treatment group, a difference of 0.5 should be significant for 80% of the simulations at a test level of 5%. The treatment means at the centers varied to provide a variety of interaction patterns, but were chosen to provide a difference between the two treatment means of 0.5 in the case of equal sample sizes across centers.

The predicted treatment means for each treatment at each center and a description of the type of interaction exhibited by the respective pattern of means for

these simulated datasets are provided in Table 6. (The patterns are ordered in the table based on the difference parameters presented below; the pattern numbers are for identification but have no numerical significance.)

Table 6

Description of Two-Center Simulation Patterns, Interaction Type, Predicted Mean Responses and Effect Sizes for Each Treatment-by-Center Cell

Pattern Number	Interaction Type	Center 1			Center 2			Effect Difference
		Trt. 1	Trt. 2	Effect	Trt. 1	Trt. 2	Effect	
1	None	0.0	0.5	0.5	0.0	0.5	0.5	0.0
9	None	-1.0	-0.5	0.5	1.0	1.5	0.5	0.0
8	Quant.	0.0	0.25	0.25	0.0	0.75	0.75	0.5
2	Quant.	0.0	0.0	0.0	0.0	1.0	1.0	1.0
6	Quant.	0.0	0.0	0.0	1.0	2.0	1.0	1.0
3	Qual.	0.0	-0.25	-0.25	0.0	1.25	1.25	1.5
4	Qual.	0.0	-0.5	-0.5	0.0	1.5	1.5	2.0
7	Qual.	0.5	0.0	-0.5	0.0	1.5	1.5	2.0
5	Qual.	0.0	-1.0	-1.0	0.0	2.0	2.0	3.0

For two patterns, 1 and 9, there was no interaction, i.e. the difference between the treatment means (effect difference) was equal at both centers. For Pattern 1, the means for each respective treatment were also equal at each center; while for Pattern 9, the mean responses for each respective treatment varied between the two centers.

Three patterns, 2, 6 and 8, were simulated to represent quantitative interaction. For Pattern 8, the treatment effect was 0.25 at Center 1 and 0.75 at Center 2. For

Patterns 2 and 6, there was no difference between treatments at Center 1 and a difference between treatments of 1.0 at Center 2. These latter two patterns show a greater degree of quantitative interaction than the first pattern. For Pattern 2, the response levels of Treatment 1 were the same at both centers. For Pattern 6, the response levels varied across centers for each treatment.

Patterns 3, 4, 5 and 7 display qualitative interaction. The patterns vary in degree of qualitative interaction; Pattern 5 has the most, Pattern 3 has the least and Patterns 4 and 7 are intermediate. Pattern 3 has a small negative treatment effect at one center and a positive effect at the other. In Patterns 4 and 7, there is a larger negative effect at one center and a larger positive effect at the other. The negative and positive affects are both larger in Pattern 5. The difference between Pattern 4 and Pattern 7 is that the response for Treatment 1 is the same at both centers for Pattern 4 and varies for Pattern 7.

5.5.2 Patterns With No Difference Between Treatments (No Overall Treatment Effect)

Two additional datasets were simulated with two treatment groups at two centers, but with no difference in overall mean response between the two treatment groups. The overall variance and the number of patients per treatment were the same as for the simulations above. Although the difference between the means of Treatment 1 and Treatment 2 was equal to zero, treatment means at the centers varied to provide a variety of interaction patterns. The predicted treatment means for each treatment at each center and a description of the type of interaction exhibited by the respective pattern of means for these simulated datasets are provided in Table 7. (The patterns are ordered in the table based on the difference parameters presented below; the pattern numbers are for identification but have no numerical significance.)

For Pattern 11, there was no interaction and the means for each respective treatment were also equal at each center.

Table 7

Description of Two-Center Simulation Patterns, Interaction Type, Predicted Mean Responses and Effect Sizes for Each Treatment-by-Center Cell

Pattern Number	Interaction Type	Center 1			Center 2			Effect Difference
		Trt. 1	Trt. 2	Effect	Trt. 1	Trt. 2	Effect	
11	None	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	Qual.	0.0	1.0	1.0	1.0	0.0	-1.0	2.0

Pattern 10 displays qualitative interaction, with the differences between the two treatment groups at each center canceling each other out in the calculation of the overall difference between treatments.

The interaction patterns simulated for the two-center trials are presented in Figure 4, Figure 5, Figure 6 and Figure 7.

5.5.3 Difference Parameters

The relationship of the sample size patterns to each other can be described by defining the patterns in terms of two difference parameters: 1) Difference between Treatment 1 and Treatment 2 at Center 1, 2) Additional difference between Treatment 1 and Treatment 2 at Center 2. These parameters are displayed in Figure 8 for the patterns described above. In this figure interactions to the right of the Y-axis are quantitative and those to the left are qualitative. The degree of qualitative interaction corresponds with a shift to the left. The patterns with no overall treatment difference

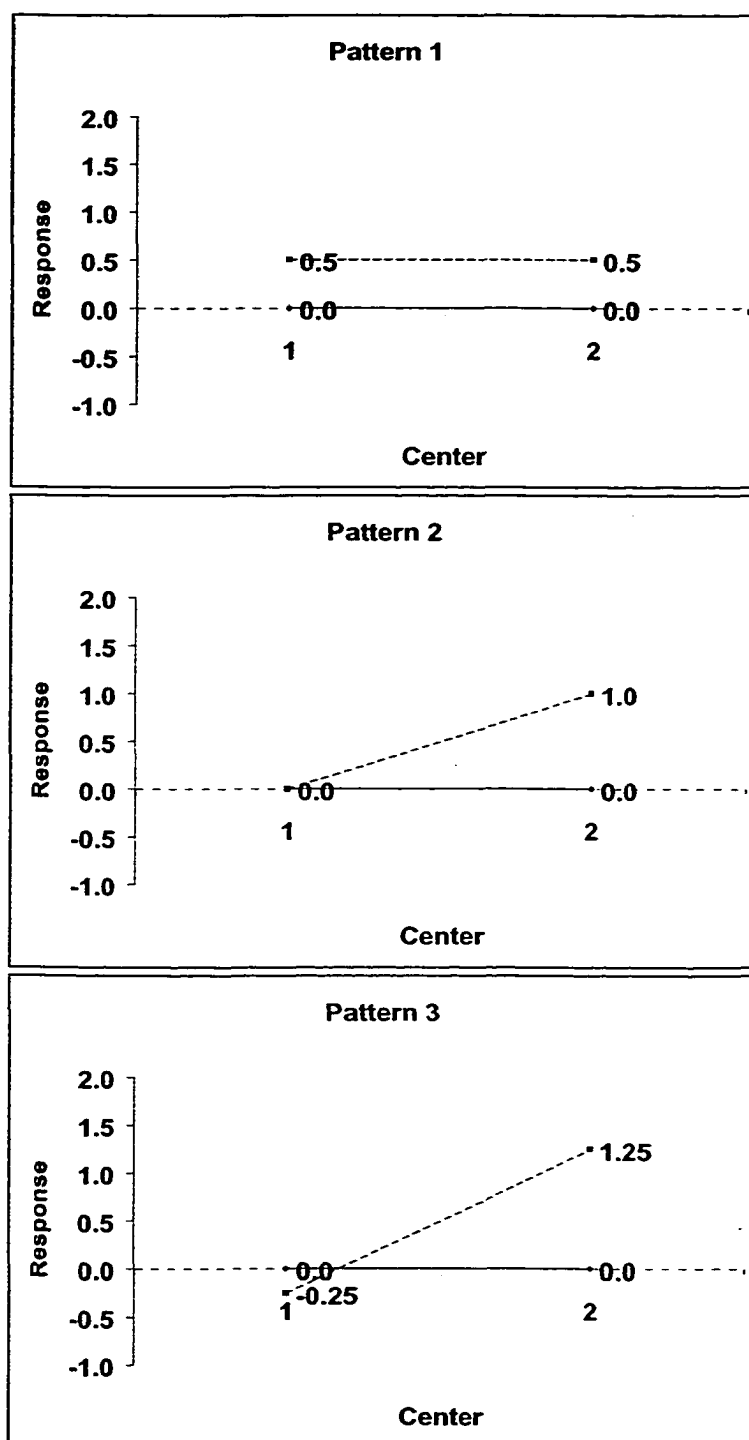


Figure 4. Simulated Interaction Patterns for Two Centers: Patterns 1 - 3 (Treatment 1 —, Treatment 2 --).

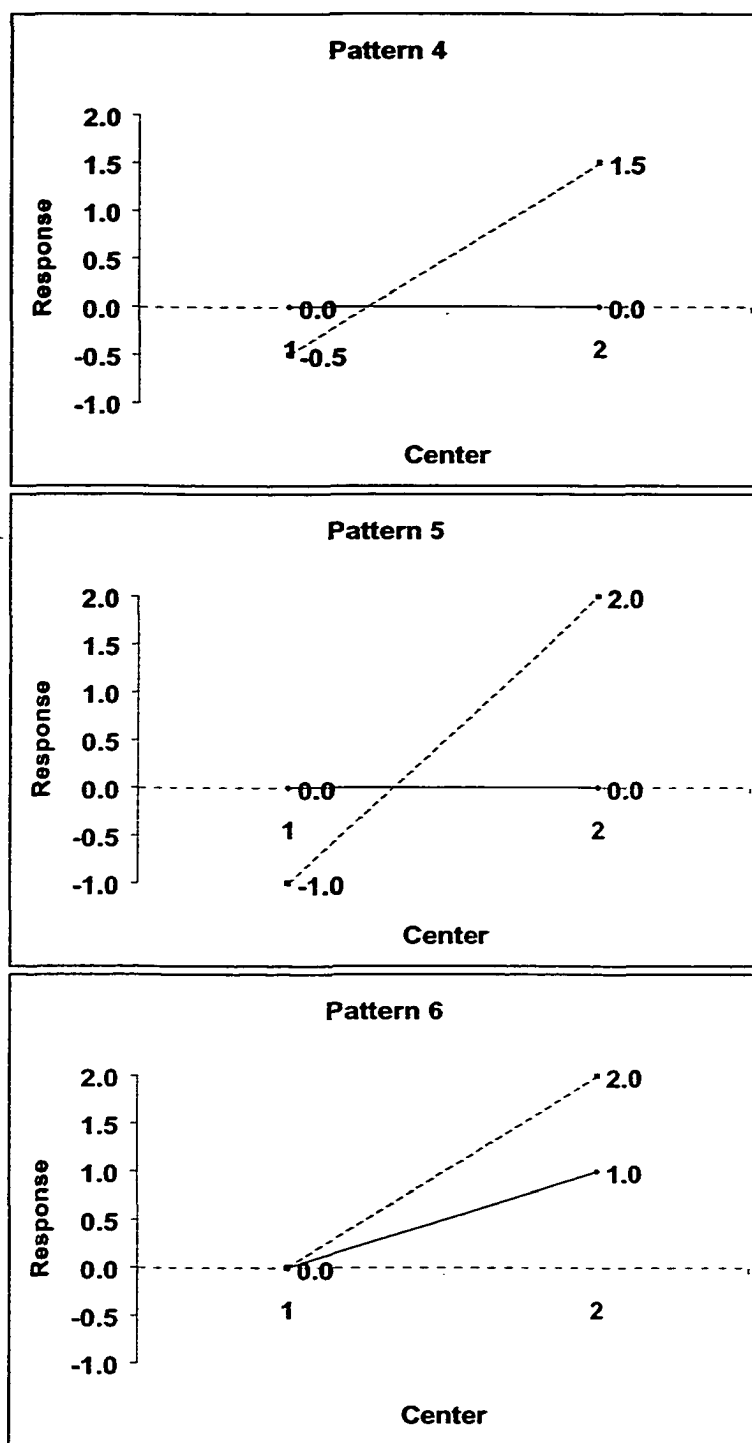


Figure 5. Simulated Interaction Patterns for Two Centers: Patterns 4 - 6 (Treatment 1 —, Treatment 2 --).

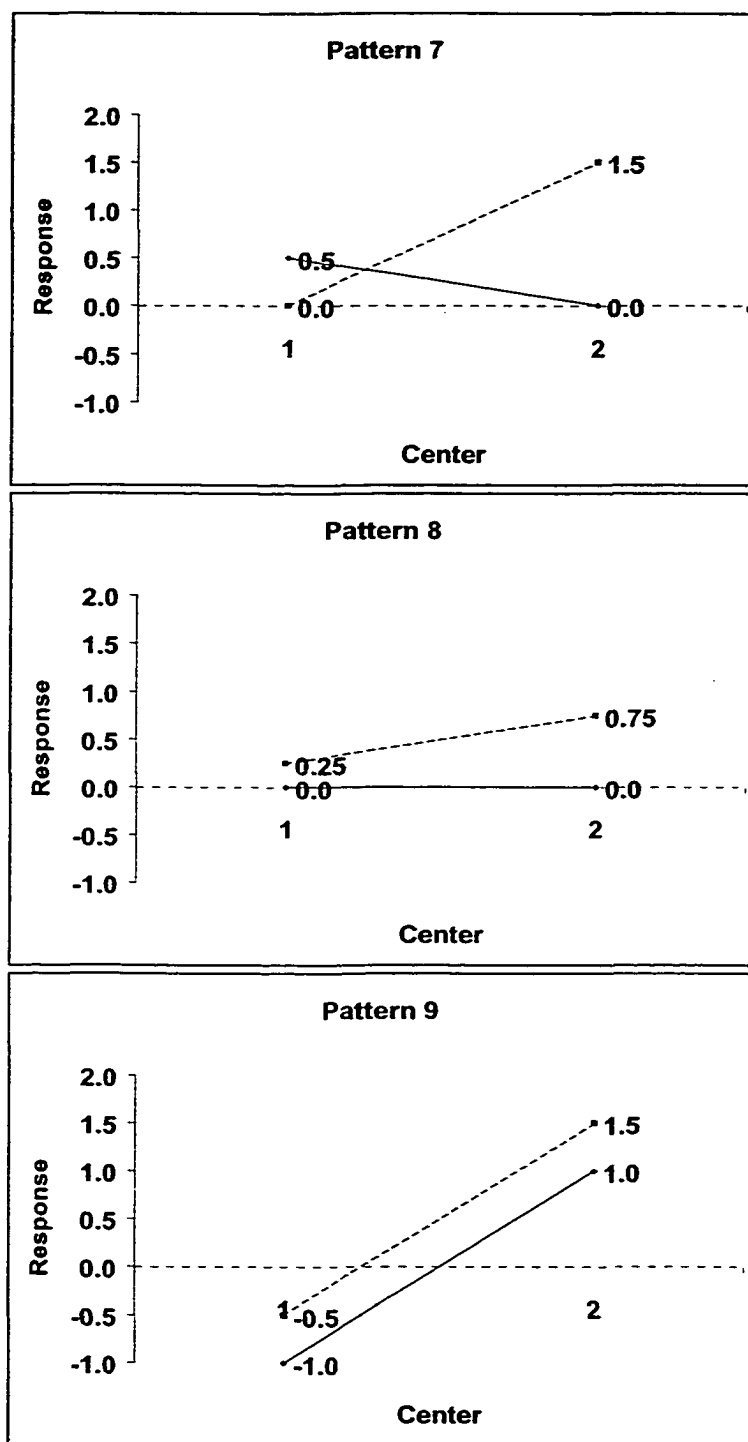


Figure 6. Simulated Interaction Patterns for Two Centers: Patterns 6 - 9 (Treatment 1 —, Treatment 2 --).

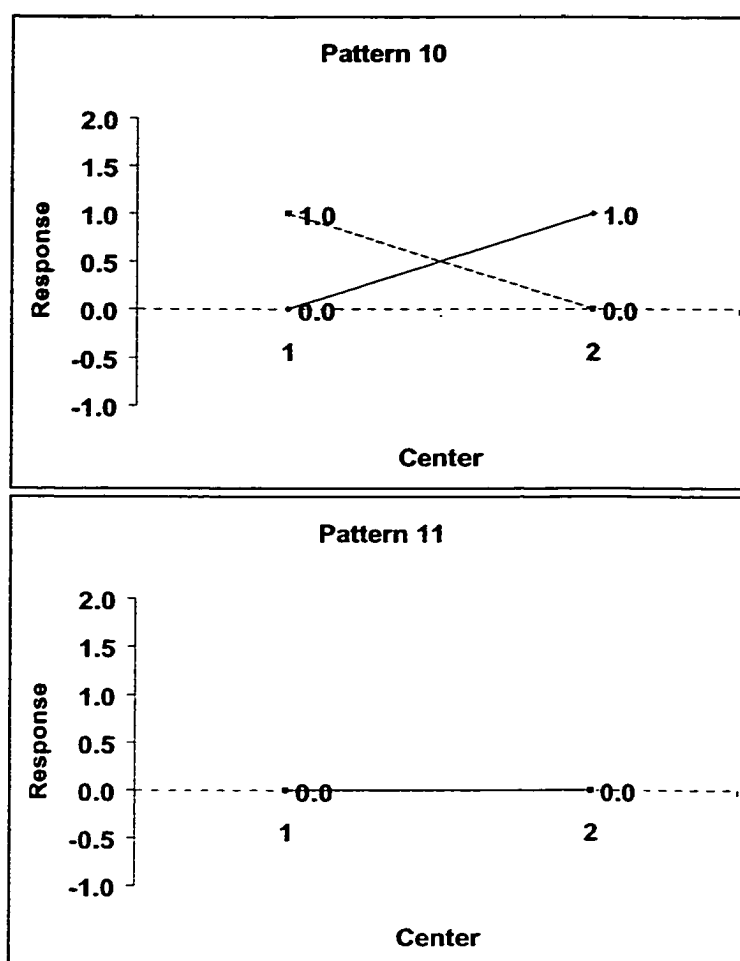


Figure 7. Simulated Interaction Patterns for Two Centers: Patterns 10 - 11
(Treatment 1 —, Treatment 2 ---).

are linearly related, as are those with an overall treatment difference of 0.5. These parameters are identical for three respective pairs of patterns: Patterns 1 and 9 (0.5, 0.0), Patterns 2 and 6 (0.0, 1.0) and Patterns 4 and 7 (-0.5, 2.0).

5.5.4 Sample Sizes

The effect of sample sizes on the results of the tested methods was examined by varying sample sizes at the centers. The sample sizes of the two treatment groups

within a center were always equal to each other. Sample size patterns for the simulations of two centers are presented in Table 8.

Each of the interaction patterns above was simulated for each of the sample size cases to provide an assessment of the combined effect of sample size and interaction pattern.

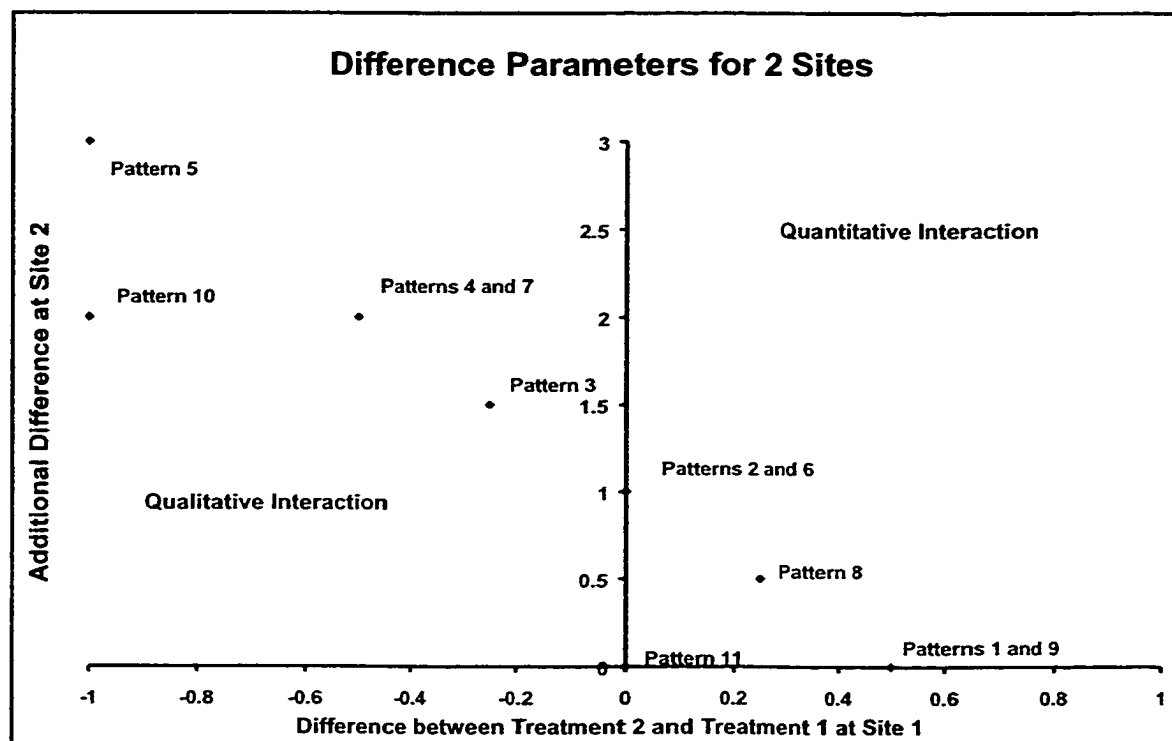


Figure 8. Difference Parameters for Interaction Patterns for Two Centers.

5.6 Simulation of Trials With Three Centers

The same methodology of comparison was extended to three centers. Datasets with two treatment groups at three centers were simulated. Eleven patterns of datasets were simulated: nine patterns with a difference between treatments (overall treatment effect) and two patterns without a difference between treatments (no overall

treatment effect). Patterns were selected to display either no interaction, quantitative interaction or qualitative interaction.

Table 8
Sample Sizes for Each Treatment Group for Simulations of Interaction Patterns for Two Centers

Case	Center 1	Center 2	Total
2.1	32	32	64
2.2	43	21	64
2.3	21	43	64
2.4	54	10	64
2.5	10	54	64

5.6 1 Patterns With a Difference Between Treatments (Overall Treatment Effect)

The total number of patients per treatment, the overall difference between the means of Treatment 1 and Treatment 2 across the three centers and the overall variance were the same as for the simulation for two centers. These parameters were determined so that the simulated data would have an overall mean treatment effect that would be significant at the 5% level in 80% of simulated datasets (i.e. power of 80%). The difference between the means of Treatment 1 and Treatment 2 was 0.5, the overall variance was 1.0 and the sample size was 64 patients per treatment. As in the simulation for two centers, with sixty-four patients per treatment group, a difference of 0.5 should be significant for 80% of the simulations at a test level of 5%. The treatment means at the centers varied to provide a variety of interaction patterns, but were determined to provide a difference between the two treatment means of 0.5 in the case of equal sample sizes across centers.

The predicted treatment means for each treatment at each center and a description of the type of interaction exhibited by the respective pattern of means for these simulated datasets are provided in Table 9. (The patterns are ordered in the table based on the magnitude of the interaction; the pattern numbers are for identification but have no numerical significance.)

Pattern 1 was simulated to provide no interaction, i.e. the difference between the treatment means was equal at all three centers. The mean responses for each respective treatment were also equal across all centers.

Four patterns, 2, 3, 4 and 5, display quantitative interaction. In Pattern 2, there are small treatment effects at two centers and a larger effect at the third. One of the centers in Pattern 3 shows no treatment effect and the other two have equal effects. All three effects in Pattern 4 are different: one is zero, one is large and the other is intermediate. In Pattern 5, there is no effect at two of the centers and a large effect at the third.

There are also four patterns, 6, 7, 8 and 9, that were simulated to show qualitative interaction. Pattern 6 has a negative effect at one center and positive effects equal to each other at the other centers. Pattern 7 has a negative effect at one center and two unequal positive effects at the other two centers. One center in Pattern 8 has a negative effect, one has a positive effect and at one center there is no difference between the responses of the two treatments. In Pattern 9, there is a small positive effect at one center, a small negative effect at one center and large positive effect at the third center.

Table 9

Description of Three-Center Simulation Patterns, Interaction Type and Predicted Mean Responses for Each Treatment-by-Center Cell

Pattern Number	Interaction Type	Center 1		Center 2		Center 3	
		Treat. 1	Treat. 2	Treat. 1	Treat. 2	Treat. 1	Treat. 2
1	No Interaction	0.0	0.5	0.0	0.5	0.0	0.5
2	Quantitative	0.0	0.2	0.0	0.2	0.0	1.1
3	Quantitative	0.0	0.0	0.0	0.75	0.0	0.75
4	Quantitative	0.0	0.0	0.0	0.5	0.0	1.0
5	Quantitative	0.0	0.0	0.0	0.0	0.0	1.5
6	Qualitative	0.0	-0.5	0.0	1.0	0.0	1.0
7	Qualitative	0.0	-0.5	0.0	0.5	0.0	1.5
8	Qualitative	0.0	-0.5	0.0	0.0	0.0	2.0
9	Qualitative	0.0	-0.25	0.0	-0.25	0.0	2.0

5.6 2 Patterns With No Difference Between Treatments (No Overall Treatment Effect)

Two additional datasets with two treatment groups at three centers were simulated, but with no difference in mean overall response between the two treatment groups. The overall variance and the number of patients per treatment were the same as for the simulations above. Although the difference between the means of Treatment 1 and Treatment 2 was equal to zero, treatment means at the centers varied to provide a variety of interaction patterns. The predicted treatment means for each treatment at each center and a description of the type of interaction exhibited by the

respective pattern of means for these simulated datasets are provided in Table 10. (The patterns are ordered in the table based on the magnitude of the interaction; the pattern numbers are for identification but have no numerical significance.)

Table 10

Description of Three-Center Simulation Patterns, Interaction Type and Predicted Mean Responses for Each Treatment-by-Center Cell

Pattern Number	Interaction Type	Center 1		Center 2		Center 3	
		Treat. 1	Treat. 2	Treat. 1	Treat. 2	Treat. 1	Treat. 2
10	No Interaction	0.0	0.0	0.0	0.0	0.0	0.0
11	Qualitative	0.0	-0.5	0.0	0.0	0.0	0.5

For Pattern 10, there was no interaction and the means for each respective treatment were also equal at each center.

Pattern 11 displays qualitative interaction; the treatment response is zero for one center, one center has a negative difference and the other center shows a positive response. The non-zero differences between the two treatment groups cancel each other out in the calculation of the overall treatment response.

The interaction patterns simulated for the three-center trials are presented in Figure 9, Figure 10, Figure 11, and Figure 12.

5.6 3 Sample Sizes

The effect of sample sizes on the results of the tested methods was examined by varying sample sizes at the centers. The sample sizes of the two treatment groups

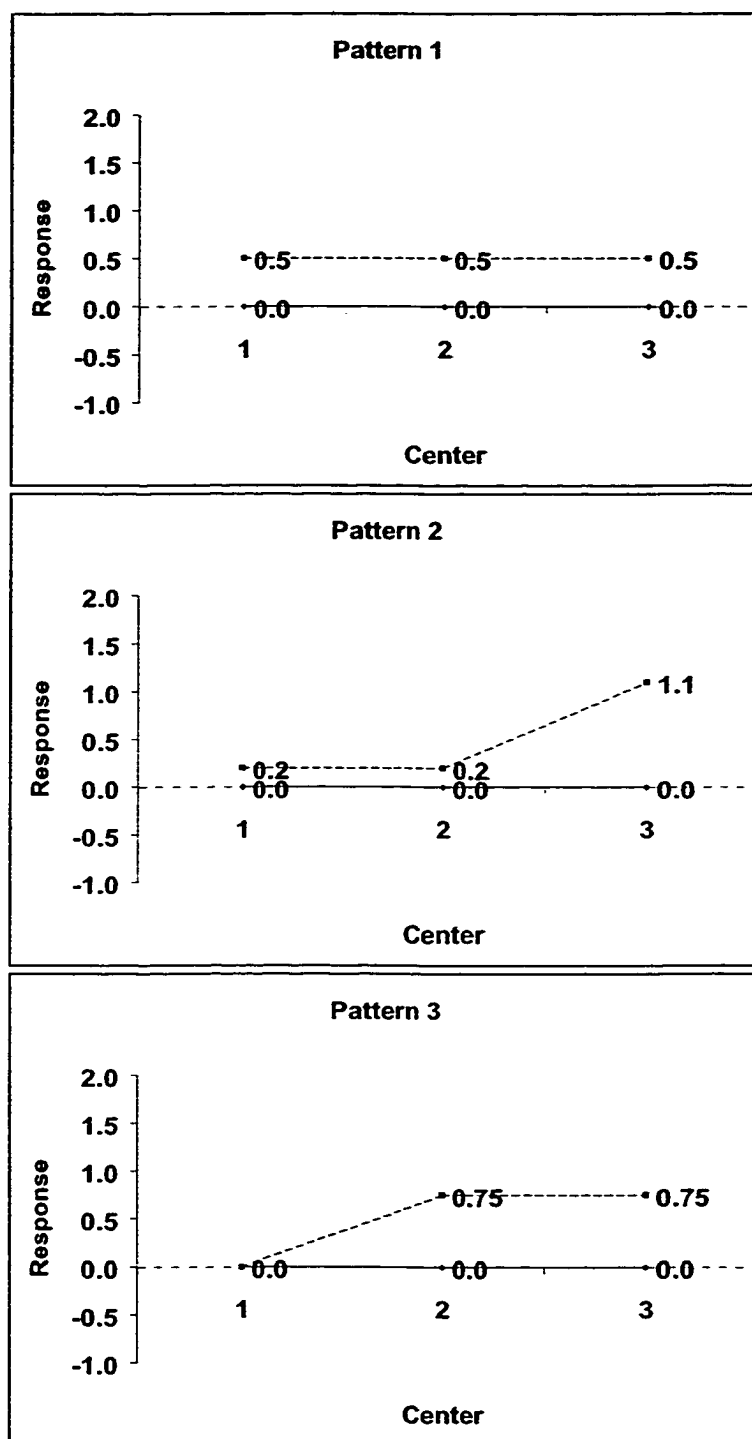


Figure 9. Simulated Interaction Patterns for Three Centers: Patterns 1 - 3 (Treatment 1 —, Treatment 2 --).

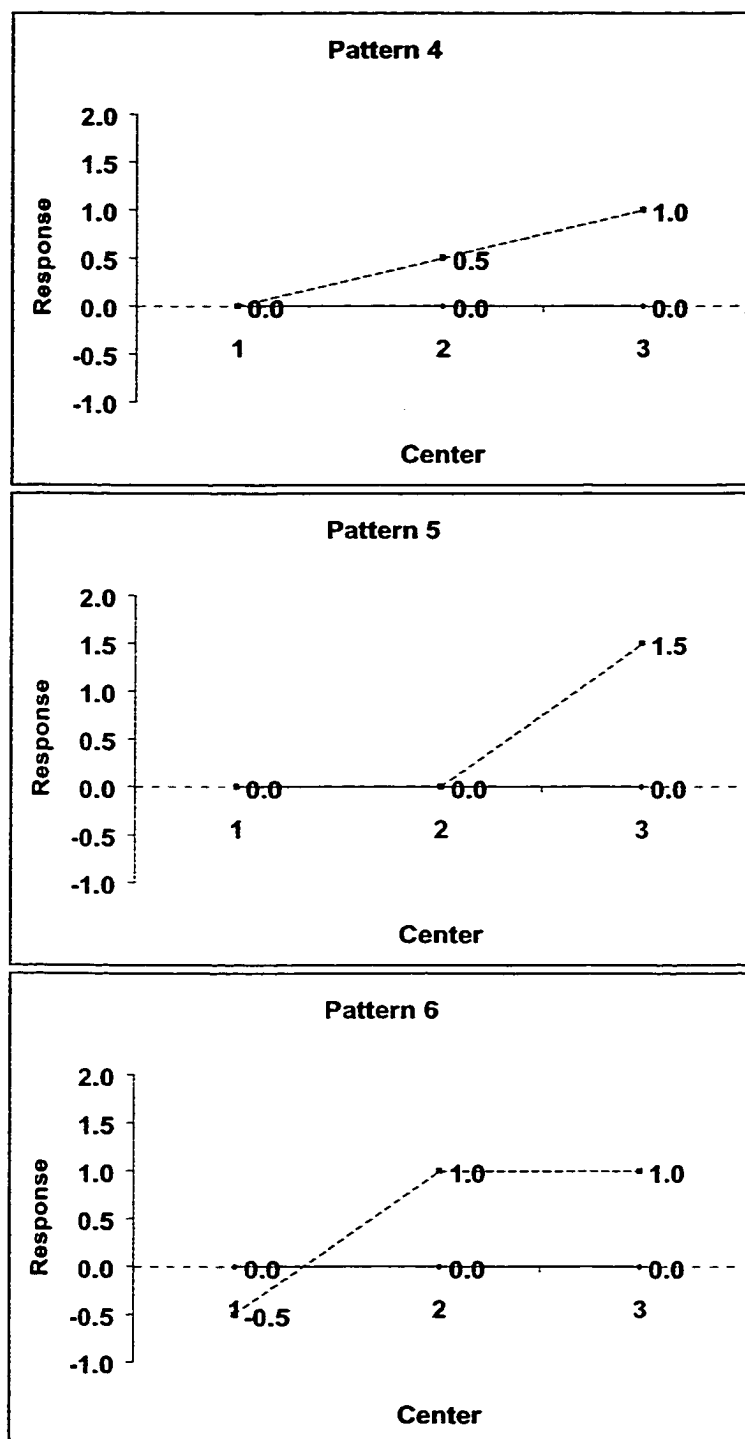


Figure 10. Simulated Interaction Patterns for Three Centers: Patterns 4 - 6 (Treatment 1 —, Treatment 2 --).

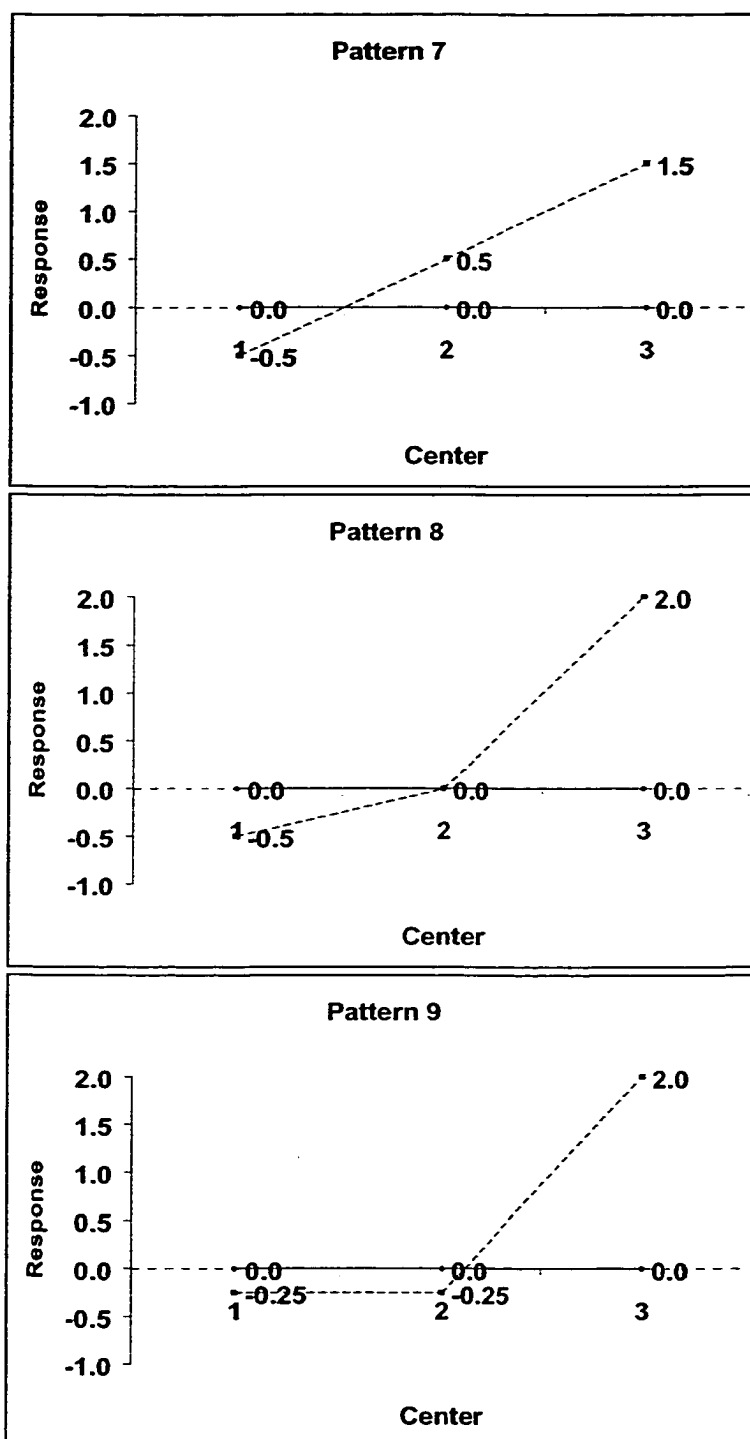


Figure 11. Simulated Interaction Patterns for Three Centers: Patterns 7 - 9 (Treatment 1 —, Treatment 2 --).

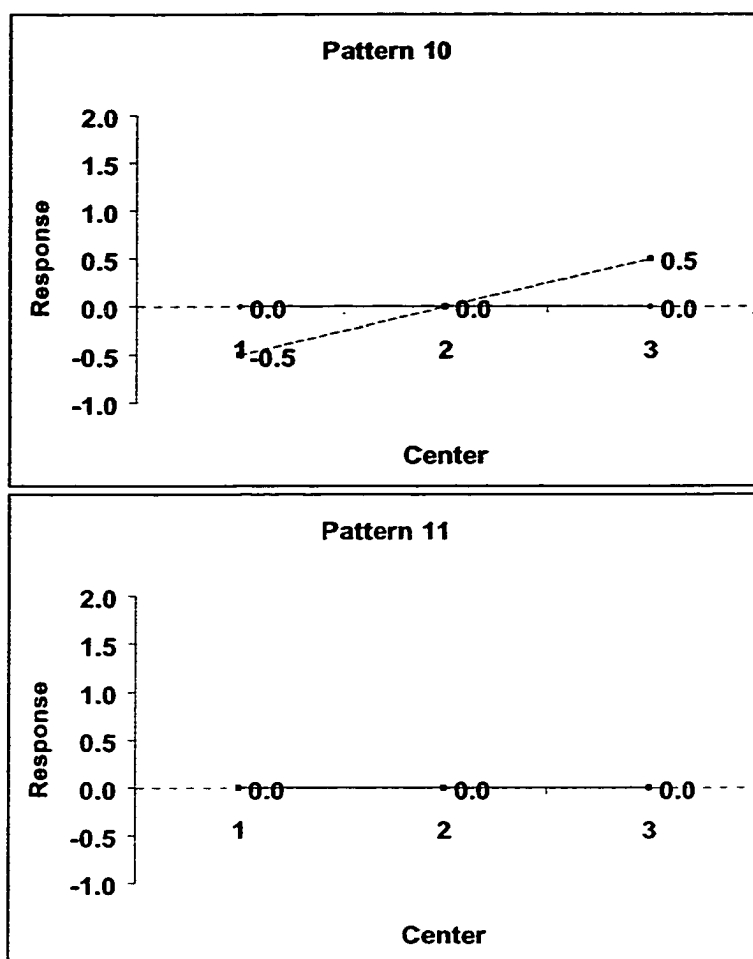


Figure 12. Simulated Interaction Patterns for Three Centers: Patterns 10 - 11 (Treatment 1 —, Treatment 2 ---).

within a center were always equal to each other. Sample size patterns for the simulations of three centers are presented in Table 11.

Each of the interaction patterns above was simulated for each of the sample size cases to provide an assessment of the combined effect of sample size and interaction patterns.

Table 11

Sample Sizes for Each Treatment Group for Simulations of Interaction Patterns
for Three Centers

Case	Center 1	Center 2	Center 3	Total
3.1	22	21	21	64
3.2	39	13	12	64
3.3	13	39	12	64
3.4	13	12	39	64
3.5	29	29	6	64
3.6	29	6	29	64
3.7	6	29	29	64

CHAPTER VI

RESULTS FROM SIMULATED DATA FOR TWO CENTERS

6.1 Introduction

The operating characteristics of the methods of identifying quantitative interaction proposed by Azzalini and Cox (1984), Gail and Simon (1985) and Ciminera et al. (1993) were examined using simulated datasets depicting a two-center trial with two treatments.

As described in Chapter V, data were simulated for 11 patterns of treatment-by-center interaction with five sample size configurations for each interaction pattern. The total sample size for each treatment group was 64, but the distribution of the patients across the two centers differed for each configuration. The sample sizes of the two treatment groups within each center were always equal to each other. For each pattern and sample size configuration, 2500 datasets were simulated and summarized.

The 2500 datasets for each pattern and sample size configuration were generated using the same initial seed number and adjusting each generated number by an appropriate amount to produce the expected treatment-by-center means. Thus differences between the patterns and configurations are not the result of any differences in the sets of random numbers.

6.2 Case 2.1: 32 Patients at Center 1 and 32 Patients at Center 2

6.2.1 General Results

The simulations for the equal sample size configuration generated 32 observations for each treatment group at each center. The average means and variances of these 2500 datasets, without any adjustment to the means for treatment and center effects, are presented in Table 12. The target value of each of these means was 0.000 and the target variance of the observations was 1.000.

Table 12

Means and Variances by Treatment and Center of Two-Center Simulated Data

Center	Treatment	Mean	Variance
1	1	-0.0004	0.973
1	2	0.0006	0.998
2	1	0.0015	1.001
2	2	-0.0066	0.984

The randomly generated numbers were consistent with the targeted values for the simulation.

6.2.2 Tests of Overall Interaction

The results of the test of significance of the treatment-by-center interaction from analysis of variance and using the H statistic proposed by Gail and Simon were obtained for each simulated dataset. The analysis of variance test was considered to be significant if the F statistic for the test had a significance level less than or equal to

10%. The H test was considered significant if the level of the associated chi-square test was less than or equal to 10%.

The percentage of significant results for each test from the 2500 simulated datasets of each interaction pattern is presented in Table 13. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

The two tests provide results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in over 99% of the individual simulations for each pattern.

The power of the tests to detect interaction in the patterns with qualitative interaction (Patterns 3, 4, 5, 7 and 10) is at or near 100%. The power to detect the stronger quantitative interaction (Patterns 2 and 6) is much greater than the power to detect the weaker quantitative interaction (Pattern 8). In the patterns with no interaction (Patterns 1, 9 and 11), the error rate for the detection of interaction differs slightly between the two methods. However, the error rates for both methods are at or near the expected error rate of 10%.

6.2.3 Methods for the Detection of Qualitative Interaction

The number of simulated datasets with qualitative interaction was evaluated using the tests of Azzalini and Cox and Gail and Simon, and the method of Ciminera et al. Both of the extended methods of Azzalini and Cox (exact and approximate methods) presented in Chapter IV were used. The methods of Gail and Simon and Ciminera et al. were evaluated using the variance estimate from the analysis of variance, as presented in Chapters III and IV.

Table 13

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 2.1

Pattern and Int. Type*	Characteristics of Simulation Patterns				Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Analysis of Variance	Gail and Simon H Statistic
1 NI	Yes	None	0.5	0.5	9	10
9 NI	Yes	None	0.5	0.5	9	10
8 QN	Yes	Quant.	0.25	0.75	39	39
2 QN	Yes	Quant.	0.0	1.0	88	88
6 QN	Yes	Quant.	0.0	1.0	88	88
3 QL	Yes	Qual.	-0.25	1.25	99	99
4 QL	Yes	Qual.	-0.5	1.5	100	100
7 QL	Yes	Qual.	-0.5	1.5	100	100
5 QL	Yes	Qual.	-1.0	2.0	100	100
11 NI	No	None	0.0	0.0	9	10
10 QL	No	Qual.	1.0	-1.0	100	100

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The significance of the Azzalini and Cox test was calculated for each simulated dataset for both the exact and approximate procedures. Each dataset was tested for qualitative interaction using Gail and Simon's test with the minimum of Q^+ and Q^- . For each of these procedures, a dataset was considered to have a significant

qualitative interaction if the level of the test was less than or equal to 10%. The number of significant tests from the 2500 simulated datasets was summarized.

The results of the method of Ciminera et al. were examined for each dataset. Any dataset with “pushed back” (destandardized) mean differences which were not all either positive or negative was considered to show “substantial evidence” of the presence of qualitative interaction. The number of datasets showing “substantial evidence” was summarized.

The results of the three methods are presented in Table 14, along with the percentages of datasets in which the treatment differences in the raw data have differing signs across the two centers. A frequently used, ad-hoc, method of evaluating data from multicenter trials for the presence of qualitative interaction is to examine the data for differing signs among the treatment differences. If all of the signs of the treatment differences are not either positive or negative, then the dataset is considered to have qualitative interaction.

Since these methods are designed to detect qualitative interaction, the percentages of qualitative interactions detected in Patterns 1, 9, 8, 2, 6 and 11 will be considered to be error rates for these tests. The percentages of qualitative interactions detected in Patterns 3, 4, 7, 5 and 10 will be considered to be the power of the tests.

As expected the exact and approximate methods of Azzalini and Cox provide equivalent results when sample sizes are equal. In the discussion of the results in this section, the two methods will not be differentiated.

All three methods provide a more realistic approach to determining the presence of important qualitative interaction than examining the presence of differing signs in the raw data. In the three patterns where there is no treatment effect at Center 1 (Patterns 2, 6 and 11), the raw data show evidence of a qualitative

interaction in approximately 50% of the simulations. This coincides with the expectation that 50% of the treatment effect values at Center 1 will be less than the mean of zero and 50% greater than zero.

Table 14

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 2.1

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	1	1	0	0	4
9 NI	1	1	0	0	4
8 QN	4	4	1	0	16
2 QN	23	23	10	5	50
6 QN	23	23	10	5	50
3 QL	60	60	39	26	84
4 QL	89	89	76	63	98
7 QL	89	89	76	63	98
5 QL	100	100	100	99	100
11 NI	11	11	2	1	51
10 QL	100	100	99	98	100

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The three methods provide similar results in patterns with an overall treatment effect and no interaction (Patterns 1, 9), minimal quantitative interaction (Pattern 8) or sizeable qualitative interaction (Pattern 5). However, test results differ markedly among the three methods for the other, intermediate, interaction patterns with an overall treatment effect (Patterns 2, 6, 3, 4, 7). The number of detected qualitative interactions is consistently highest for the test of Azzalini and Cox, consistently lowest for the method of Ciminera et al. and the test of Gail and Simon is between the other two.

For Patterns 2 and 6, which are at the border of the regions of quantitative and qualitative interaction, the error rate for the test of Gail and Simon is at the appropriate level (10%). For this pattern, the test of Azzalini and Cox rejects the null hypothesis of no qualitative interaction 23% of the time, while the method of Ciminera et al. finds “substantial evidence” of qualitative interaction only 5% of the time. For Pattern 3, the mean treatment effect at Center 1 is -0.25, one-half the magnitude of the overall threshold for significance of 0.5, and the mean treatment effect at Center 2 is 1.25, two and one-half times the threshold. Positive indications of qualitative interaction for this pattern range from 26% to 60%. In Patterns 4 and 7, the mean treatment effect at Center 1 is -0.5, the magnitude of the overall threshold for significance, and the mean treatment effect at Center 2 is 1.5, which is three times the threshold. The method of Ciminera et al. identified qualitative interaction in 63% of the simulations, the test of Azzalini and Cox was significant in 89% of the simulations and the test of Gail and Simon was intermediate at 76%.

In the pattern with no interaction, and no treatment effect (Pattern 11), the error rate of the Azzalini and Cox method is highest (11%), but seems appropriate for

a test with a level of 10%. The error rates of the other two tests are much less, 1% and 2%.

6.2.4 Test of Non-inferiority

The null hypotheses that each respective treatment group is at least as good as the other treatment group at every center were tested using the test of non-inferiority proposed by Gail and Simon. The method was evaluated using the variance estimate from the analysis of variance, as described in Chapters III and IV.

Each treatment group in each dataset was tested for non-inferiority against the other using Gail and Simon's test with both Q^+ and Q^- . A dataset was considered to reject the null hypothesis that the mean of Treatment 1 is equal to or greater than the mean of Treatment 2 (or vice versa) if the level of the test was less than or equal to 2.5%. The number of significant tests from the 2500 simulated datasets was summarized.

The results of the methods are presented in Table 15.

In the pattern where the mean of each treatment group is equal to the other at each center (Pattern 11), the error rate of the test is at the appropriate error level, 2.5% (rounded to 3%). For Pattern 10, there is no overall treatment effect and each treatment is numerically, if not statistically, superior to the other at one of the two centers. The power of the test to show that neither treatment is equal to or superior to the other at both centers is 95%. This is similar to Pattern 5, where the overall treatment effect is significant but there is a sizeable qualitative interaction.

Pattern 11 is the only pattern where, for at least one of the centers, Treatment 2 is not numerically superior to Treatment 1 by at least 0.5, which represents the clinically relevant difference the study is powered to detect. For all patterns other than

Pattern 11, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 is at least 73%. For Patterns 1, 9, 8, 2, and 6 the overall treatment effect is equal to the clinically relevant difference (with Treatment 2 being superior) and

Table 15

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers for Case 2.1

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at Both Centers	Treatment 1 is not \geq Treatment 2 at Both Centers
1 NI	0	73
9 NI	0	73
8 QN	0	82
2 QN	1	96
6 QN	1	96
3 QL	9	100
4 QL	36	100
7 QL	36	100
5 QL	94	100
11 NI	3	3
10 QL	95	95

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

Treatment 2 is equal to or greater than Treatment 1 at both centers, i.e. there is either no interaction or quantitative interaction. For these patterns, the error rate of not

finding Treatment 2 to be equal to or greater than Treatment 1 at both centers is less than or equal to 1%. For Pattern 3, where there is a qualitative interaction with the negative treatment effect, at Center 1, of one-half of the clinically relevant difference, the rejection rate of the test that Treatment 2 is equal to or greater than Treatment 1 at both centers is 9%. For the other patterns with qualitative interaction (i.e. where Treatment 2 is not greater than or equal to Treatment 1 at both centers, Patterns 4, 7 and 5) the power is 36%, 36% and 94%.

6.2.5 Two-stage Test Results

The purpose of two-stage testing is to improve the power and lower the error rate for the test of the overall treatment difference by using a preliminary test to select the appropriate final analysis model. The Type II model is selected as the final model when there is no consequential evidence that the interaction term needs to be included in the model. The existence of such evidence justifies the inclusion of the interaction term in the model and the Type III model is selected as the final analysis model.

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 16.

For all patterns, the Type III and Type II overall treatment means for both treatments are near the targeted values of the simulation. This is the expected result for the case of equal sample sizes. For the Type III model, the tests of significance of the difference between the two means are consistent with the selected power and error rate for the simulation. The power to detect differences, when they are present (all patterns except Patterns 11 and 10), is near 80%; and the error rate, when there is no between-treatment difference (Patterns 11 and 10), is near 5%. The results of the Type II model are similar to those of the Type III model when there is no interaction

Table 16

**Treatment Means for Each Treatment Calculated From Type II and Type III
Analyses and the Percentages of Simulations With Significant Tests
That Treatment 2 is Not Equal to Treatment 1 for Case 2.1**

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III	Type II	Type III	Type II	Type III	Type II
	Model	Model				
1 NI	80	80	0.001	0.001	0.497	0.497
9 NI	80	80	0.001	0.001	0.497	0.497
8 QN	80	79	0.001	0.001	0.497	0.497
2 QN	80	78	0.001	0.001	0.497	0.497
6 QN	80	78	0.501	0.501	0.997	0.997
3 QL	80	75	0.001	0.001	0.497	0.497
4 QL	80	72	0.001	0.001	0.497	0.497
7 QL	80	72	0.251	0.251	0.747	0.747
5 QL	80	63	0.001	0.001	0.497	0.497
11 NI	5	5	0.001	0.001	-0.003	-0.003
10 QL	5	3	0.501	0.501	0.497	0.497

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

or minimal quantitative interaction. However, as the importance of the interaction increases, the power of the Type II model decreases. In the context of analysis of variance, this results from pooling the sum of squares due to interaction into the error sum of squares, which inflates the variance used to test the significance of the

treatment difference. For Pattern 5, the most extreme case of qualitative interaction, the power is less than 63% with a Type II model.

Table 17 summarizes test results when a two-stage testing procedure is used. The final analysis model contains the interaction term (Type III model) only if the

Table 17

Percentage of Simulations With Significant Test That Treatment 2
is Not Equal to Treatment 1 After Model Selection With
a Preliminary Test of Interaction for Case 2.1

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	80	80	80	80	80	80	80
9 NI	80	80	80	80	80	80	80
8 QN	79	79	79	79	79	79	79
2 QN	80	80	79	79	78	78	79
6 QN	80	80	79	79	78	78	79
3 QL	80	80	79	79	79	78	79
4 QL	80	80	80	80	79	79	80
7 QL	80	80	80	80	79	79	80
5 QL	80	80	80	80	80	80	80
11 NI	5	5	5	5	5	5	5
10 QL	5	5	5	5	5	5	5

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

preliminary test indicates that there is significant interaction (Analysis of Variance, Gail and Simon H statistic), significant qualitative interaction (Azzalini and Cox [both methods], Gail and Simon) or substantial evidence of qualitative interaction (Ciminera et. al, Raw Data). Otherwise, the Type II model is the final analysis model.

The percentages of differences found to be significant do not differ substantially as a result of the method used for preliminary testing. For those patterns with significant treatment effects (all patterns except Patterns 11 and 10), the power to detect the difference ranges from 78% to 80%. For Patterns 11 and 10, the error rate is 5% for all methods.

These results integrate the results for the case with equal sample sizes, presented in this section. The power of the Type III model to detect the treatment effect, when it is present, is always near 80%. The power of the Type II model is also near 80% when there is little evidence of interaction, but decreases as the amount of interaction increases. The power of the methods to detect interaction varies slightly, but is generally high in those patterns (Patterns 3, 7 and 5) where there is a treatment effect and substantial qualitative interaction. In these patterns, the more powerful Type III model is chosen.

For both Patterns 11 and 10, the error rates are close to the expected 5% error rate. The results for Pattern 10, a pattern without a treatment effect but with notable interaction, reflect those of the Type III model which is selected in more than 98% of the simulations by all methods.

6.2.6 Evaluation of Redundant Patterns

From the results above it is evident that there are three pairs of patterns in which each member of the pair gives identical results for all methods: Patterns 1 and 9, Patterns 2 and 6, and Patterns 4 and 7. These patterns were also identified in Chapter V to have identical difference parameters for each pair. In each of these pairs of patterns, the respective treatment effects at each center are the same for both members of the pair. For example, the treatment effect at Center 1 for Patterns 1 and 9 is 0.5, and the treatment effect at Center 2 for both patterns is 0.5. The difference between the two members of each pair is the mean response at each center. For example, the means at Center 1 are 0.25 and -0.75 , respectively, for Patterns 1 and 9, and the Center 2 means for the two patterns are 0.25 and 1.25, respectively. The similarity between the two members of each pair can be explained in the context of analysis of variance. In the presence of constant treatment effects at each center, the treatment by center interaction is not affected by a change in the magnitudes of the mean responses at each center.

Since the presentation and discussion of the redundant patterns do not provide additional insight, Patterns 6, 7 and 9 will not be mentioned further.

6.3 Case 2.2: 43 Patients at Center 1 and 21 Patients at Center 2

The simulations for this unequal sample size configuration generated 43 observations for each treatment group at Center 1 and 21 observations for each treatment group at Center 2. The results presented are based on 2500 simulated datasets of each interaction pattern

6.3.1 Tests of Overall Interaction

The results of the test of significance of the treatment-by-center interaction from analysis of variance and using the H statistic proposed by Gail and Simon are presented in Table 18. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

The two tests provide results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in 99% or more of the individual simulations for each pattern.

Table 18

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 2.2

Pattern and Int. Type*	Characteristics of Simulation Patterns				Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Analysis of Variance	Gail and Simon H Statistic
1 NI	Yes	None	0.5	0.5	10	10
8 QN	Yes	Quant.	0.25	0.75	36	36
2 QN	Yes	Quant.	0.0	1.0	83	84
3 QL	Yes	Qual.	-0.25	1.25	99	99
4 QL	Yes	Qual.	-0.5	1.5	100	100
5 QL	Yes	Qual.	-1.0	2.0	100	100
11 NI	No	None	0.0	0.0	10	10
10 QL	No	Qual.	1.0	-1.0	100	100

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The power of the tests to detect interaction in the patterns of qualitative interaction (Patterns 3, 4, 5 and 10) is at or near 100%. The power to detect the stronger quantitative interaction (Pattern 2) is much greater than the power to detect the weaker quantitative interaction (Pattern 8). In the patterns with no interaction (Patterns 1 and 11), the error rate for both methods is at the expected error rate of 10%.

6.3.2 Methods for the Detection of Qualitative Interaction

The percentages of simulated datasets with qualitative interaction, as evaluated using the tests of Azzalini and Cox (exact and approximate methods) and Gail and Simon, and the method of Ciminera et al., are presented in Table 19, along with the percentages of datasets with differing signs in the raw data.

The results from the exact and approximate approaches of Azzalini and Cox are slightly different in this case. The exact approach has a slightly higher rejection rate than the approximate approach.

For patterns with significant overall treatment effects, the three methods provide similar results in patterns of no interaction (Pattern 1) and sizeable qualitative interaction (Pattern 5). However, test results differ markedly among the three for the other, intermediate patterns. The number of detected qualitative interactions is consistently highest for the method of Ciminera et al., consistently lowest for the test of Gail and Simon and the test of Azzalini and Cox is between the other two. The presence of differing signs in the raw data generally has a higher rejection rate than any of the three proposed methods.

Table 19

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 2.2

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	1	1	0	1	6
8 QN	2	2	1	10	12
2 QN	22	18	10	47	49
3 QL	66	59	44	88	88
4 QL	95	93	85	99	99
5 QL	100	100	100	100	100
11 NI	10	8	2	10	51
10 QL	100	100	98	97	100

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For Pattern 2, which is at the border of the regions of quantitative and qualitative interaction, the error rate for the test of Gail and Simon is at the appropriate level (10%). For this pattern, the test of Azzalini and Cox rejects the null hypothesis of no qualitative interaction 18% or 22% of the time, while the method of Ciminera et al. finds “substantial evidence” of qualitative interaction 47% of the time. For Pattern 3, the mean treatment effect at Center 1 is -0.25, one-half the magnitude of the overall threshold for significance of 0.5, and the mean treatment effect at Center 2 is 1.25, two and one-half times the threshold. Positive indications of qualitative interaction for this pattern range from 44% to 88%. In Pattern 4, the mean

treatment effect at Center 1 is -0.5, the magnitude of the overall threshold for significance, and the mean treatment effect at Center 2 is 1.5, which is three times the threshold. The test of Gail and Simon was significant in 85% of the simulations, the test of Azzalini and Cox was significant in 95% and 93% of the simulations and the method of Ciminera et al. identified qualitative interaction in 99% of the simulations.

In the pattern with no interaction, and no treatment effect (Pattern 11), the error rates of the Azzalini and Cox and Ciminera et al. methods are highest (8 to 10%), but seem appropriate for a test with a level of 10%. The error rate of the Gail and Simon test is much less (2%).

6.3.3 Test of Non-inferiority

The percentages of significant tests of the null hypotheses that each respective treatment group is at least as good as the other treatment group at every center, tested using the test of non-inferiority proposed by Gail and Simon, are presented in Table 20.

In the pattern where the mean of each treatment group is equal to the other at each center (Pattern 11), the error rate of the test is near the appropriate error level, 2.5%. For Pattern 10, there is no overall treatment effect and each treatment is numerically, if not statistically, superior to the other at one of the two centers. The power of the test to show that neither treatment is equal to or superior to the other at both centers is highest for this pattern and for Pattern 5, where the overall treatment effect is significant but there is a sizeable qualitative interaction. For Pattern 10, the results of the test reflect the larger sample size at Center 1, where the mean of Treatment 1 is less than the mean of Treatment 2.

Table 20

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers for Case 2.2

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at Both Centers	Treatment 1 is not \geq Treatment 2 at Both Centers
1 NI	0	75
8 QN	0	69
2 QN	1	82
3 QL	12	95
4 QL	48	99
5 QL	99	100
11 NI	3	2
10 QL	82	99

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

Pattern 11 is the only pattern where, for at least one of the centers, Treatment 2 is not numerically superior to Treatment 1 by at least 0.5, which represents the clinically relevant difference the study is powered to detect. For all patterns other than Pattern 11, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 is at least 69%. For Patterns 1, 8 and 2 the overall treatment effect is equal to the clinically relevant difference (with Treatment 2 being superior) and Treatment 2 is equal to or greater than Treatment 1 at both centers, i.e. there is either no interaction or quantitative interaction. For these patterns, the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at both centers is less than or equal to 1%. For Pattern 3, where there is a qualitative interaction with the

negative treatment effect, at Center 1, of one-half of the clinically relevant difference, the rejection rate of the test that Treatment 2 is equal to or greater than Treatment 1 at both centers is 12%. For the other patterns with qualitative interaction (i.e. where Treatment 2 is not greater than or equal to Treatment 1 at both centers, Patterns 4 and 5) the power is 48% and 99%.

6.3.4 Two-stage Test Results

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 21.

There are notable differences in the means from the two models, especially for Treatment 2. The only difference between the Type III and Type II means for Treatment 1 is for Pattern 10; whereas for Treatment 2, only the means for Patterns 1 and 11 do not differ substantially. The smaller than expected Type II means for Patterns 8, 2, 3, 4, 5 reflect the relatively larger sample size at Center 1, where the Treatment 2 means are less than the respective Treatment 2 means at Center 2.

The tests of significance of the difference between the two treatment means are consistent with the calculated treatment means for the two models. The test results for the two models are only consistent for Patterns 1 and 11. For Patterns 8, 2, 3, 4 and 5, the patterns with interaction and treatment differences, the number of significant test results is higher for the Type III model than for the Type II model. For Pattern 10, with interaction but no treatment difference (in the Type III model), the number of significant tests is much higher for the Type II model than for the Type III model. The results of the test for this pattern reflect the larger sample size at Center 1, where the mean of Treatment 1 is less than the mean of Treatment 2.

Table 21

Treatment Means for Each Treatment Calculated From Type II and Type III
Analyses and the Percentages of Simulations With Significant Tests
That Treatment 2 is Not Equal to Treatment 1 for Case 2.2

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III	Type II				
	Model	Model	Type III	Type II	Type III	Type II
1 NI	75	80	0.000	0.000	0.496	0.497
8 QN	75	63	0.000	0.000	0.496	0.412
2 QN	75	42	0.000	0.000	0.496	0.326
3 QL	75	23	0.000	0.000	0.496	0.240
4 QL	75	9	0.000	0.000	0.496	0.154
5 QL	75	2	0.000	0.000	0.496	-0.018
11 NI	5	4	0.000	0.000	-0.004	-0.003
10 QL	5	39	0.500	0.328	0.496	0.669

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

Table 22 summarizes test results when a two-stage testing procedure is used to select the final analysis model.

For Patterns 1, 5, 11, 10, the percentages of differences found to be significant did not differ substantially depending on the method used as a preliminary test. Patterns 1 and 5 have significant treatment effects and no interaction (Pattern 1) or substantial qualitative interaction (Pattern 5). Both have power greater than 75% for all methods. For other patterns with a significant treatment effect (all patterns except

Patterns 11 and 10), the power to detect the difference ranged from 46% to 75%. For Patterns 11 and 10, the error rate was near the expected error rate of 5% for all methods.

Table 22

Percentage of Simulations With Significant Test That Treatment 2
is Not Equal to Treatment 1 After Model Selection With
a Preliminary Test of Interaction for Case 2.2

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	79	79	80	80	80	80	79
8 QN	72	72	64	64	63	67	67
2 QN	73	73	52	49	46	66	66
3 QL	75	75	63	58	48	73	73
4 QL	75	75	73	72	67	75	75
5 QL	75	75	75	75	75	75	75
11 NI	5	5	4	4	4	5	5
10 QL	5	5	5	5	6	6	5

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

Except for Pattern 1, the power is highest using the two tests of overall interaction as preliminary testing procedures. Of the qualitative interaction detection methods, the method of Ciminera et al. consistently provides the highest power and the test of Gail and Simon has the lowest power. Pre-testing with these methods

generally produced the lowest numbers of significant tests for Patterns 2 and 3, patterns with quantitative or minimal qualitative interaction.

6.4 Case 2.3: 21 Patients at Center 1 and 43 Patients at Center 2

The simulations for this unequal sample size configuration generated 21 observations for each treatment group at Center 1 and 43 observations for each treatment group at Center 2.

6.4.1 Tests of Overall Interaction

The results of the test of significance of the treatment-by-center interaction from analysis of variance and using the H statistic proposed by Gail and Simon are presented in Table 23. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

The two tests provide results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in 99% or more of the individual simulations for each pattern.

The power of the tests to detect interaction in the patterns of qualitative interaction (Patterns 3, 4, 5 and 10) is at or near 100%. The power to detect the stronger quantitative interaction (Pattern 2) is much greater than the power to detect the weaker quantitative interaction (Pattern 8). In the patterns with no interaction (Patterns 1 and 11), the error rate for both methods is near the expected error rate of 10%.

Table 23

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 2.3

Pattern and Int. Type*	Characteristics of Simulation Patterns				Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Analysis of Variance	Gail and Simon H Statistic
1 NI	Yes	None	0.5	0.5	8	8
8 QN	Yes	Quant.	0.25	0.75	37	37
2 QN	Yes	Quant.	0.0	1.0	85	85
3 QL	Yes	Qual.	-0.25	1.25	99	99
4 QL	Yes	Qual.	-0.5	1.5	100	100
5 QL	Yes	Qual.	-1.0	2.0	100	100
11 NI	No	None	0.0	0.0	8	8
10 QL	No	Qual.	1.0	-1.0	100	100

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

6.4.2 Methods for the Detection of Qualitative Interaction

The percentages of simulated datasets with qualitative interaction, as evaluated using the tests of Azzalini and Cox (exact and approximate methods) and Gail and Simon, and the method of Ciminera et al., are presented in Table 24, along with the percentages of datasets with differing signs in the raw data.

The results from the exact and approximate approaches of Azzalini and Cox are slightly different in this case. The exact approach has a slightly lower rejection

rate than the approximate approach for patterns with moderate interaction and a treatment difference (Patterns 8, 2, 3, 4).

Table 24

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 2.3

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	1	1	0	1	6
8 QN	6	7	1	1	21
2 QN	23	27	10	9	50
3 QL	52	57	33	30	80
4 QL	82	85	63	61	95
5 QL	100	100	98	98	100
11 NI	10	9	2	7	50
10 QL	100	100	98	97	100

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For patterns with significant overall treatment effects, the three methods provide similar results in patterns with no interaction (Pattern 1) and sizeable qualitative interaction (Pattern 5). However, test results differ markedly among the three for the other, intermediate patterns. The number of detected qualitative interactions is consistently higher for the test of Azzalini and Cox (both procedures)

than for the other two methods. The presence of differing signs in the raw data is generally higher than the rejection rates of any of the three proposed methods.

For Pattern 2, which is at the border of the regions of quantitative and qualitative interaction, the error rate for the test of Gail and Simon and method of Ciminera et al. is at or near the appropriate level of 10%. For this pattern, the test of Azzalini and Cox rejects the null hypothesis of no qualitative interaction 23% or 27% of the time. For Pattern 3, positive indications of qualitative interaction range from 30% to 57%. In Pattern 4, the test of Gail and Simon was significant in 63% of the simulations, the test of Azzalini and Cox was significant in 82% and 85% of the simulations and the method of Ciminera et al. identified qualitative interaction in 61% of the simulations.

In the pattern with no interaction, and no treatment effect (Pattern 11), the error rates of the Azzalini and Cox and Ciminera et al. methods are highest (7 to 10%), but seem appropriate for a test with a level of 10%. The error rate of the Gail and Simon test is much less (2%).

6.4.3 Test of Non-inferiority

The percentages of significant tests of the null hypotheses that each respective treatment group is at least as good as the other treatment group at every center, tested using the test of non-inferiority proposed by Gail and Simon, are presented in Table 25.

In the pattern where the mean of each treatment group is equal to the other at each center (Pattern 11), the error rate of the test is at the appropriate error level, 2.5%. The power of the test to show that neither treatment is equal to or superior to the other at both centers is highest for Patterns 10 and 5, patterns with sizeable

qualitative interactions. For Pattern 10, the results of the test reflect the larger sample size at Center 2, where the mean of Treatment 2 is less than the mean of Treatment 1. For Pattern 5, the results of the test reflect the smaller sample size at Center 1, where the mean of Treatment 2 is less than the mean of Treatment 1.

Table 25

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers for Case 2.3

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at Both Centers	Treatment 1 is not \geq Treatment 2 at Both Centers
1 NI	0	73
8 QN	0	90
2 QN	1	99
3 QL	6	100
4 QL	24	100
5 QL	83	100
11 NI	3	3
10 QL	99	80

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For all patterns other than Pattern 11, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 is at least 73%. For Patterns 1, 8 and 2, the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at both centers is less than or equal to 1%. For Pattern 3, the rejection rate of the test that Treatment 2 is equal to or greater than Treatment 1 at both centers is 6%. For the

other patterns with qualitative interaction (i.e. where Treatment 2 is not greater than or equal to Treatment 1 at both centers, Patterns 4 and 5) the power is 24% and 83%.

6.4.4 Two-stage Test Results

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 26.

Table 26

Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 2.3

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III	Type II	Type III	Type II	Type III	Type II
	Model	Model				
1 NI	74	79	0.001	0.001	0.497	0.497
8 QN	74	90	0.001	0.001	0.497	0.583
2 QN	74	96	0.001	0.001	0.497	0.669
3 QL	74	98	0.001	0.001	0.497	0.754
4 QL	74	99	0.001	0.001	0.497	0.840
5 QL	74	100	0.001	0.001	0.497	1.012
11 NI	4	5	0.001	0.001	-0.003	-0.003
10 QL	4	41	0.501	0.673	0.497	0.325

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

There are notable differences in the means from the two models, especially for Treatment 2. The only difference between the Type III and Type II means for Treatment 1 is for Pattern 10; whereas for Treatment 2, only the means for Patterns 1 and 11 do not differ substantially. The larger than expected Type II means for Patterns 8, 2, 3, 4, 5 reflect the relatively larger sample size at Center 2, where the Treatment 2 means are greater than the respective Treatment 2 means at Center 1.

The tests of significance of the difference between the two treatment means are consistent with the calculated treatment means for the two models. The test results for the two models are only consistent for Patterns 1 and 11. For Patterns 8, 2, 3, 4 and 5, the patterns with interaction and treatment differences, the number of significant test results is higher for the Type II model than for the Type III model. For Pattern 10, with interaction but no treatment difference (in the Type III model), the number of significant tests is much higher for the Type II model than for the Type III model. The results of the test reflect the larger sample size at Center 2, where the mean of Treatment 2 is less than the mean of Treatment 1.

Table 27 summarizes test results when a two-stage testing procedure is used to select the final analysis model.

For Patterns 1, 5, 11, 10, the percentages of differences found to be significant did not differ substantially depending on the method used as a preliminary test. Patterns 1 and 5 have significant treatment effects and no interaction (Pattern 1) or substantial qualitative interaction (Pattern 5). Both have power greater than 74% for all methods. For other patterns with a significant treatment effect (all patterns except Patterns 11 and 10), the power to detect the difference ranged from 74% to 90%. For Patterns 11 and 10, the error rate was near the expected error rate of 5% for all methods.

Table 27

**Percentage of Simulations With Significant Test That Treatment 2
is Not Equal to Treatment 1 After Model Selection With
a Preliminary Test of Interaction for Case 2.3**

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	78	79	79	79	79	79	78
8 QN	80	80	87	86	89	89	80
2 QN	75	75	82	80	89	90	75
3 QL	74	74	75	75	80	80	74
4 QL	74	74	74	74	75	75	74
5 QL	74	74	74	74	74	74	74
11 NI	5	5	5	5	5	5	5
10 QL	4	4	5	5	5	6	5

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For all patterns, the number of significant tests is higher using the qualitative interaction detection methods than using the two tests of overall interaction as preliminary testing procedures. With the exception of Patterns 2 and 3, the results of the qualitative interaction detection methods were very consistent.

6.5 Comparison of Case 2.2 and Case 2.3

In both Cases 2.2 and 2.3, one of the two centers has twice the patients per treatment group as the other. The distribution of patients between centers is inversely

proportional in the two cases. In Case 2.2, the larger numbers of patients (43 per treatment group) are at Center 1 and the smaller numbers (21 per group) are at Center 2. In Case 2.3, the smaller numbers of patients (21 per treatment group) are at Center 1 and the larger numbers (43 per group) are at Center 2. Although the sample size has no practical effect on the Type III treatment means or their differences for any of the patterns, the Type II treatment means and their differences are substantially affected for some patterns. The differences between the treatment means for Case 2.2 and Case 2.3 are presented in Table 28.

The differences in the Type II means are a direct result of the treatment means defined by the patterns. For all patterns except Pattern 10, the Treatment 1 means are near 0.0 for both centers. The Treatment 2 means for Patterns 1 and 11 are the same at both centers. For Patterns 8, 2, 3, 4, and 5, the means for Treatment 2 at Center 2 are positive and are larger in absolute value than the means at Center 1. Hence, for these patterns the Type II means for Treatment 2 are larger in Case 2.3 because of the larger sample size at Center 2. For Pattern 10, the Treatment 1 and 2 means are 0 and 1 and 1 and 0 at Centers 1 and 2, respectively.

The difference in the number of patients per treatment group between the two cases also has an important effect on the results of the tests for qualitative interaction. The differences between Cases 2.2 and 2.3 are presented in Table 29, along with a count of the differences in the number of datasets with differing signs in the raw data.

From the raw data, we see that the difference between the cases in the percentage of simulations with sign reversals is less than 10% for all patterns, and is 1% or less for five of the eight. In the patterns with no interaction or substantial qualitative interaction (Patterns 1, 11, 5 and 10), none of methods of detecting qualitative interaction show important differences between the two cases. The largest

differences between the two cases are for Patterns 2, 3 and 4 and the method showing the most substantial differences between the two cases was the method of Ciminera et al.

Table 28
Differences (Case 2.2 – Case 2.3) in Treatment Means
for Type III and Type II Models

Pattern and Int. Type*	Differences in Treatment 1		Differences in Treatment 2	
	Means		Means	
	Type III	Type II	Type III	Type II
1 NI	-0.001	-0.001	-0.001	0.001
8 QN	-0.001	-0.001	-0.001	-0.171
2 QN	-0.001	-0.001	-0.001	-0.343
3 QL	-0.001	-0.001	-0.001	-0.515
4 QL	-0.001	-0.001	-0.001	-0.687
5 QL	-0.001	-0.001	-0.001	-1.030
11 NI	-0.001	-0.001	-0.001	0.001
10 QL	-0.001	-0.345	-0.001	0.345

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

A further comparison of the two cases is a summary of the results of the two-stage testing procedure. Table 30 summarizes the difference in test results of the two cases when a two-stage testing procedure is used.

Table 29

Differences (Case 2.2 – Case 2.3) in Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction.

Pattern and Int. Type*	Difference in Percentages With Significant Tests			Difference in Percentages With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	0	0	0	0	0
8 QN	-4	-5	0	9	-9
2 QN	-1	-9	0	38	-1
3 QL	14	2	11	58	8
4 QL	13	8	22	38	4
5 QL	0	0	2	2	0
11 NI	0	-1	0	3	1
10 QL	0	0	0	0	0

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The tests of overall interaction are minimally affected by the difference in sample sizes portrayed in the two cases. The only patterns where the tests show differences greater than 1% are Patterns 8 and 2, which have quantitative interaction. In the patterns with no interaction or substantial qualitative interaction (patterns 1, 11, 5, and 10), all of the methods would give similar results if used as a preliminary test in a two-stage testing procedure. For the patterns with interaction, but little or no qualitative interaction (Patterns 8, 2 and 3), the sample size shift between the two

centers has an important effect. For Pattern 4, only the test of Gail and Simon is substantially affected.

Table 30

Difference (Case 2.2 – Case 2.3) in Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	1	0	1	1	1	1	1
8 QN	-9	-8	-23	-23	-26	-22	-13
2 QN	-3	-2	-30	-31	-43	-24	-9
3 QL	1	1	-12	-17	-32	-7	-1
4 QL	1	1	-1	-2	-7	0	1
5 QL	1	1	1	1	1	1	1
11 NI	0	0	0	0	0	0	0
10 QL	0	0	0	0	1	0	0

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

6.6 Case 2.4: 54 Patients at Center 1 and 10 Patients at Center 2

The 2500 simulations for this unequal sample size configuration each generated 54 observations for each treatment group at Center 1 and 10 observations for each treatment group at Center 2.

6.6.1 Tests of Overall Interaction

The results of the test of significance of the treatment-by-center interaction from analysis of variance and using the H statistic proposed by Gail and Simon are presented in Table 31. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

Table 31

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 2.4

Pattern and Int. Type*	Characteristics of Simulation Patterns				Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Analysis of Variance	Gail and Simon H Statistic
1 NI	Yes	None	0.5	0.5	9	9
8 QN	Yes	Quant.	0.25	0.75	25	25
2 QN	Yes	Quant.	0.0	1.0	65	66
3 QL	Yes	Qual.	-0.25	1.25	91	92
4 QL	Yes	Qual.	-0.5	1.5	99	99
5 QL	Yes	Qual.	-1.0	2.0	100	100
11 NI	No	None	0.0	0.0	9	9
10 QL	No	Qual.	1.0	-1.0	99	99

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The two tests provide results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in 99% or more of the individual simulations for each pattern.

The power of the tests to detect interaction in the patterns with qualitative interaction (Patterns 4, 5 and 10) is at or near 100%. The power to detect the stronger quantitative interaction (Pattern 2) is much greater than the power to detect the weaker quantitative interaction (Pattern 8). In the patterns with no interaction (Patterns 1 and 11), the error rate for both methods is near the expected error rate of 10%.

6.6.2 Methods for the Detection of Qualitative Interaction

The percentages of simulated datasets with qualitative interaction, as evaluated using the tests of Azzalini and Cox (exact and approximate methods) and Gail and Simon, and the method of Ciminera et al., are presented in Table 32, along with the percentages of datasets with differing signs in the raw data.

The results from the exact and approximate approaches of Azzalini and Cox differ in this case. The exact approach has a higher rejection rate than the approximate approach, except for Patterns 1, 5 and 10.

For patterns with significant overall treatment effects, the three methods provide similar results in patterns with no interaction (Pattern 1) and in the case of sizeable qualitative interaction (Pattern 5). However, test results differ markedly among the three for the other, intermediate patterns. The number of detected qualitative interactions is consistently highest for the method of Ciminera et al. and is consistently lowest for the test of Gail and Simon and the approximate method of Azzalini and Cox. Results for the exact method of Azzalini and Cox are intermediate.

The presence of differing signs in the raw data is higher than the rejection rate of any of the three proposed methods.

Table 32

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 2.4

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	3	5	1	2	15
8 QN	2	1	0	7	14
2 QN	21	9	8	44	49
3 QL	70	48	47	88	90
4 QL	97	90	89	99	100
5 QL	100	100	100	100	100
11 NI	10	4	2	17	50
10 QL	93	96	84	91	99

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For Pattern 2, which is at the border of the regions of quantitative and qualitative interaction, the error rates for the test of Gail and Simon and the approximate method of Azzalini and Cox are near the appropriate level (10%). For this pattern, the exact test of Azzalini and Cox rejects the null hypothesis of no qualitative interaction 21% of the time, while the method of Ciminera et al. finds “substantial evidence” of qualitative interaction 44% of the time. For Pattern 3, the

mean treatment effect at Center 1 is -0.25, one-half the magnitude of the overall threshold for significance of 0.5, and the mean treatment effect at Center 2 is 1.25, two and one-half times the threshold. Positive indications of qualitative interaction for this pattern range from 47% to 88%. In Pattern 4, the mean treatment effect at Center 1 is -0.5, the magnitude of the overall threshold for significance, and the mean treatment effect at Center 2 is 1.5, which is three times the threshold. The test of Gail and Simon was significant in 89% of the simulations, the test of Azzalini and Cox was significant in 97% and 90% of the simulations and the method of Ciminera et al. identified qualitative interaction in 99% of the simulations.

In the pattern with no interaction, and no treatment effect (Pattern 11), the error rates of the Azzalini and Cox and Ciminera et al. methods are highest (4, 10 and 17%), with the exact test of Azzalini and Cox having the appropriate level of 10%. The error rate of the Gail and Simon test is much less (2%).

6.6.3 Test of Non-inferiority

The percentages of significant tests of the null hypotheses that each respective treatment group is at least as good as the other treatment group at every center, tested using the test of non-inferiority proposed by Gail and Simon, are presented in Table 33.

In the pattern where the mean of each treatment group is equal to the other at each center (Pattern 11), the error rate of the test is at the appropriate error level, 2.5%. For Pattern 10, there is no overall treatment effect and each treatment is numerically, if not statistically, superior to the other at one of the two centers. The power of the test to show that neither treatment is equal to or superior to the other at both centers is highest for this pattern and for Patterns 4 and 5, where the overall

treatment effect is significant but there are sizeable qualitative interactions. For Pattern 10, the results of the test reflect the larger sample size at Center 1, where the mean of Treatment 1 is less than the mean of Treatment 2.

Table 33

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers for Case 2.4

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at Both Centers	Treatment 1 is not \geq Treatment 2 at Both Centers
1 NI	0	75
8 QN	0	47
2 QN	1	48
3 QL	14	66
4 QL	60	83
5 QL	100	98
11 NI	3	3
10 QL	47	100

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

Pattern 11 is the only pattern where, for at least one of the centers, Treatment 2 is not numerically superior to Treatment 1 by at least 0.5, which represents the clinically relevant difference the study is powered to detect. For all patterns other than Pattern 11, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 ranges from 47% to 98%. For all of these patterns other than Patterns 8 and 2, which have quantitative interaction, the power is at least 66%. For Patterns 1,

8 and 2 the overall treatment effect is equal to the clinically relevant difference (with Treatment 2 being superior) and Treatment 2 is equal to or greater than Treatment 1 at both centers, i.e. there is either no interaction or quantitative interaction. For these patterns, the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at both centers is less than or equal to 1%. For Pattern 3, where there is a qualitative interaction with the negative treatment effect, at Center 1, of one-half of the clinically relevant difference, the rejection rate of the test that Treatment 2 is equal to or greater than Treatment 1 at both centers is 14%. For the other patterns with qualitative interaction (i.e. where Treatment 2 is not greater than or equal to Treatment 1 at both centers, Patterns 4 and 5) the power is 60% and 100%.

6.6.4 Two-stage Test Results

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 34.

There are notable differences in the means from the two models, especially for Treatment 2. The only difference between the Type III and Type II means for Treatment 1 is for Pattern 10; whereas for Treatment 2, only the means for Patterns 1 and 11 do not differ substantially. The smaller than expected Type II means for Patterns 8, 2, 3, 4, 5 reflect the relatively larger sample size at Center 1, where the Treatment 2 means are less than the respective Treatment 2 means at Center 2.

The numbers of significant tests of difference between the two treatment means are generally much lower than the expected 80% for both models. The results of the tests of significance are consistent with the calculated treatment means for the two models. The test results for the two models are only consistent for Pattern 11. For Patterns 8, 2, 3 and 4, patterns with interaction and treatment differences, the

number of significant test results is higher for the Type III model than for the Type II model. For Pattern 10, with interaction but no treatment difference (in the Type III model), the number of significant tests is much higher for the Type II model than for the Type III model. The results of the test for this pattern reflect the larger sample size at Center 1, where the mean of Treatment 1 is less than the mean of Treatment 2.

Table 34

Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 2.4

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III	Type II	Type III	Type II	Type III	Type II
	Model	Model				
1 NI	52	80	0.000	0.000	0.491	0.497
8 QN	52	44	0.000	0.000	0.491	0.326
2 QN	52	12	0.000	0.000	0.491	0.154
3 QL	52	4	0.000	0.000	0.491	-0.018
4 QL	52	15	0.000	0.000	0.491	-0.190
5 QL	52	78	0.000	0.000	0.491	-0.534
11 NI	5	5	0.000	0.000	-0.009	-0.003
10 QL	5	96	0.500	0.156	0.491	0.841

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

Table 35 summarizes test results when a two-stage testing procedure is used to select the final analysis model.

Table 35

Percentage of Simulations With Significant Test That Treatment 2
is Not Equal to Treatment 1 After Model Selection With
a Preliminary Test of Interaction for Case 2.4

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	77	77	79	78	80	79	71
8 QN	55	55	44	44	44	46	45
2 QN	49	49	19	14	14	32	31
3 QL	52	52	34	21	23	47	46
4 QL	52	52	50	45	46	52	52
5 QL	52	52	52	52	52	52	52
11 NI	7	7	5	5	5	6	5
10 QL	5	5	9	7	18	12	5

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For Patterns 1, 5 and 11, the percentages of differences found to be significant did not differ substantially depending on the method used as a preliminary test. Patterns 1 and 5 have significant treatment effects and no interaction (Pattern 1) or substantial qualitative interaction (Pattern 5). Pattern 1 has power greater than 77% for all methods, whereas Pattern 5 had a 52% rejection rate for all tests. For other

patterns with a significant treatment effect (all patterns except Patterns 11 and 10), the power to detect the difference ranged from 14% to 55%. For Pattern 11, the error rate was near the expected error rate of 5% for all methods. The error rate for Pattern 10 ranged from 5 to 18%.

Except for Pattern 1, the power is highest using the two tests of overall interaction as preliminary testing procedures, although for Patterns 4 and 5 other methods achieve the same power. Of the qualitative interaction detection methods, the method of Ciminera et al. generally provides the highest power. Pre-testing with these methods generally produced the lowest numbers of significant tests for Patterns 2 and 3.

6.7 Case 2.5: 10 Patients at Center 1 and 54 Patients at Center 2

The simulations for this unequal sample size configuration generated 10 observations for each treatment group at Center 1 and 54 observations for each treatment group at Center 2.

6.7.1 Tests of Overall Interaction

The results of the test of significance of the treatment-by-center interaction from analysis of variance and using the H statistic proposed by Gail and Simon are presented in Table 36. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

The two tests provide results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in 99% or more of the individual simulations for each pattern.

Table 36

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 2.5

Pattern and Int. Type*	Characteristics of Simulation Patterns				Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Analysis of Variance	Gail and Simon H Statistic
1 NI	Yes	None	0.5	0.5	11	11
8 QN	Yes	Quant.	0.25	0.75	27	28
2 QN	Yes	Quant.	0.0	1.0	65	66
3 QL	Yes	Qual.	-0.25	1.25	92	92
4 QL	Yes	Qual.	-0.5	1.5	99	99
5 QL	Yes	Qual.	-1.0	2.0	100	100
11 NI	No	None	0.0	0.0	11	11
10 QL	No	Qual.	1.0	-1.0	99	99

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The power of the tests to detect interaction in the patterns with qualitative interaction (Patterns 4, 5 and 10) is at or near 100%. The power to detect the stronger quantitative interaction (Pattern 2) is much greater than the power to detect the weaker quantitative interaction (Pattern 8). In the patterns with no interaction (Patterns 1 and 11), the error rate for both methods is near the expected error rate of 10%.

6.7.2 Methods for the Detection of Qualitative Interaction

The percentages of simulated datasets with qualitative interaction, as evaluated using the tests of Azzalini and Cox (exact and approximate methods) and Gail and Simon, and the method of Ciminera et al., are presented in Table 37, along with the percentages of datasets with differing signs in the raw data.

Table 37

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 2.5

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	4	5	1	3	14
8 QN	10	13	5	8	28
2 QN	23	28	11	17	50
3 QL	44	49	24	35	70
4 QL	64	70	44	56	86
5 QL	92	95	82	89	99
11 NI	9	5	2	17	48
10 QL	92	94	82	89	98

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The results from the exact and approximate approaches of Azzalini and Cox are slightly different in this case. The exact approach has a slightly lower rejection rate than the approximate approach for all patterns except Pattern 11.

The three methods provide similar results only for Pattern 1, which has a treatment difference, but no interaction. However, test results differ notably among the three for all other patterns. The number of detected qualitative interactions is consistently higher for both methods of Azzalini and Cox than for the other two methods, with the sole exception of Pattern 11. The presence of differing signs among the treatment differences in the raw data is always higher than the rejection rates of any of the three proposed methods.

For Pattern 2, the error rate for the test of Gail and Simon is near the appropriate level (10%). For this pattern, the test of Azzalini and Cox rejects the null hypothesis of no qualitative interaction 23% or 28% of the time and the test of Ciminera et al. has a 17% rejection rate. For Pattern 3, positive indications of qualitative interaction range from 24% to 49%. In Pattern 4, the test of Gail and Simon was significant in 44% of the simulations, the test of Azzalini and Cox was significant in 64% and 70% of the simulations and the method of Ciminera et al. identified qualitative interaction in 56% of the simulations.

In the pattern with no interaction, and no treatment effect (Pattern 11), the error rates of the Azzalini and Cox and Ciminera et al. methods are highest (5, 7 and 17%), with the exact method of Azzalini and Cox near the appropriate test level of 10%. The error rate of the Gail and Simon test is less (2%).

6.7.3 Test of Non-inferiority

The percentages of significant tests of the null hypotheses that each respective treatment group is at least as good as the other treatment group at every center, tested using the test of non-inferiority proposed by Gail and Simon, are presented in Table 38.

Table 38

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at Both Centers for Case 2.5

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at Both Centers	Treatment 1 is not \geq Treatment 2 at Both Centers
1 NI	0	74
8 QN	0	95
2 QN	2	100
3 QL	5	100
4 QL	12	100
5 QL	46	100
11 NI	4	2
10 QL	100	46

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

In the pattern where the mean of each treatment group is equal to the other at each center (Pattern 11), the error rate of the test is near the appropriate error level,

2.5%. The results of this test for Pattern 10 reflect the larger sample size at Center 2, where the mean of Treatment 2 is less than the mean of Treatment 1.

For all patterns other than Patterns 10 and 11; the power to detect that Treatment 1 is not equal to or greater than Treatment 2 is at least 74%. For Patterns 1, 8 and 2, the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at both centers is less than or equal to 2%. For Pattern 3, the rejection rate of the test that Treatment 2 is equal to or greater than Treatment 1 at both centers is 5%. For the other patterns with qualitative interaction (i.e. where Treatment 2 is not greater than or equal to Treatment 1 at both centers, Patterns 4 and 5) the power is 12% and 46%.

6.7.4 Two-stage Test Results

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 39.

There are notable differences in the means from the two models, especially for Treatment 2. The only difference between the Type III and Type II means for Treatment 1 is for Pattern 10; whereas for Treatment 2, only the means for Patterns 1 and 11 do not differ substantially. The larger than expected Type II means for Patterns 8, 2, 3, 4, 5 reflect the relatively larger sample size at Center 2, where the Treatment 2 means are greater than the respective Treatment 2 means at Center 1.

The tests of significance of the difference between the two treatment means are generally consistent with the calculated treatment means for the two models. The test results for the two models are relatively consistent only for Pattern 11. For Patterns 8, 2, 3, 4 and 5, the patterns with interaction and treatment differences, the number of significant test results is higher for the Type II model than for the Type III

model. This result is also true for Pattern 1, which has a treatment difference but no interaction. For Pattern 10, with interaction but no treatment difference (in the Type III model), the number of significant tests is much higher for the Type II model than for the Type III model. The results of the test reflect the larger sample size at Center 2, where the mean of Treatment 2 is less than the mean of Treatment 1.

Table 39

Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 2.5

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III	Type II	Type III	Type II	Type III	Type II
	Model	Model				
1 NI	53	79	0.001	0.000	0.499	0.497
8 QN	53	96	0.001	0.000	0.499	0.669
2 QN	53	100	0.001	0.000	0.499	0.841
3 QL	53	100	0.001	0.000	0.499	1.013
4 QL	53	100	0.001	0.000	0.499	1.185
5 QL	53	100	0.001	0.000	0.499	1.529
11 NI	6	5	0.001	0.000	-0.001	-0.003
10 QL	6	96	0.501	0.844	0.499	0.154

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

Table 40 summarizes test results when a two-stage testing procedure is used to select the final analysis model.

Table 40

Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 2.4

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	76	76	78	77	79	78	72
8 QN	75	74	87	84	92	89	71
2 QN	58	58	77	73	89	83	58
3 QL	53	53	61	58	77	67	53
4 QL	53	53	54	53	61	56	53
5 QL	53	53	53	53	53	53	53
11 NI	8	8	5	5	5	6	6
10 QL	7	7	11	9	20	13	6

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For Patterns 1, 5, 11, the percentages of differences found to be significant did not differ substantially depending on the method used as a preliminary test. Patterns 1 and 5 have significant treatment effects and no interaction (Pattern 1) or substantial qualitative interaction (Pattern 5). The power for Pattern 1 ranges from 76% to 79% and the power for Pattern 5 is consistent at 53% for all methods. For other patterns

with a significant treatment effect (all other patterns except Patterns 11 and 10), the power to detect the difference ranged from 53% to 92%. For Pattern 11, the error rate was near the expected error rate of 5% for all methods and for Pattern 10, the error rate ranged from 7% (for the tests of overall interaction) to 20% (when Gail and Simon was used as a pretest).

For all patterns except Patterns 4 and 5, the number of significant tests is higher using the qualitative interaction detection methods than using the two tests of overall interaction as preliminary testing procedures. With the exception of Patterns 2 and 3, the results of the qualitative interaction detection methods were consistent.

6.8 Comparison of Case 2.4 and Case 2.5

In both Cases 2.4 and 2.5, one of the two centers has over 5 times the number of patients per treatment group of the other. The distribution of patients between centers is inversely proportional in the two cases. In Case 2.4, the larger numbers of patients (54 per treatment group) are at Center 1 and the smaller numbers (10 per group) are at Center 2. In Case 2.5, the smaller numbers of patients (10 per treatment group) are at Center 1 and the larger numbers (54 per group) are at Center 2. Although the sample size has no practical effect on the Type III treatment means or their differences for any of the patterns, the Type II treatment means and their differences are substantially affected for some patterns. The differences between the treatment means for Case 2.4 and Case 2.5 are presented in Table 41.

The differences in the Type II means are a direct result of the treatment means defined by the patterns. For all patterns except Pattern 10, the Treatment 1 means are near 0.0 for both centers. The Treatment 2 means for Patterns 1 and 11 are the same at both centers. For Patterns 8, 2, 3, 4, and 5, the means for Treatment 2 means at

Center 2 are positive and are larger in absolute value than the means at Center 1. Hence, for these patterns the Type II means for Treatment 2 are larger in Case 2.5 because of the larger sample size at Center 2. For Pattern 10, the Treatment 1 and 2 means are 0 and 1 and 1 and 0 at Centers 1 and 2, respectively.

Table 41
Differences (Case 2.4 – Case 2.5) in Treatment Means
for Type III and Type II Models

Pattern and Int. Type*	Differences in Treatment 1		Differences in Treatment 2	
	Means		Means	
	Type III	Type II	Type III	Type II
1 NI	-0.001	0.000	-0.008	0.000
8 QN	-0.001	0.000	-0.008	-0.344
2 QN	-0.001	0.000	-0.008	-0.688
3 QL	-0.001	0.000	-0.008	-1.031
4 QL	-0.001	0.000	-0.008	-1.375
5 QL	-0.001	0.000	-0.008	-2.063
11 NI	-0.001	0.000	-0.008	0.000
10 QL	-0.001	-0.687	-0.008	0.687

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The difference in the number of patients per treatment group between the two cases also has an important effect on the results of the tests for qualitative interaction. The differences between Cases 2.4 and 2.5 are presented in Table 42, along with a count of the differences in the number of datasets with differing signs in the raw data.

Table 42

Differences (Case 2.4 – Case 2.5) in Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction.

Pattern and Int. Type*	Difference in Percentages With Significant Tests			Difference in Percentages With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	-1	0	0	-1	1
8 QN	-8	-12	-5	-1	-14
2 QN	-2	-19	-3	27	-1
3 QL	26	-1	23	53	20
4 QL	33	20	45	43	14
5 QL	8	5	18	11	1
11 NI	1	-1	0	0	2
10 QL	1	2	2	2	1

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

From the raw data, we see that the difference between the cases in the percentage of simulations with sign reversals is less than 20% for all patterns, and is 2% or less for five of the eight. In the patterns with no interaction or substantial qualitative interaction (Patterns 1, 11 and 10), none of methods of detecting qualitative interaction show important differences between the two cases. The largest differences between the two cases are for Patterns 2, 3 and 4 and the method showing the most substantial differences between the two cases was the method of Ciminera et al.

A further comparison of the two cases is a summary of the results of the two-stage testing procedure. Table 43 summarizes the difference in test results of the two cases when a two-stage testing procedure is used.

Table 43

Differences (Case 2.4 – Case 2.5) in Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	1	0	1	1	1	1	-1
8 QN	-20	-20	-42	-40	-48	-43	-26
2 QN	-9	-9	-58	-59	-75	-51	-27
3 QL	-1	-1	-27	-38	-54	-21	-7
4 QL	-1	-1	-4	-8	-15	-4	-1
5 QL	-1	-1	-1	-1	-1	-1	-1
11 NI	-1	-1	0	0	0	-1	-1
10 QL	-1	-1	-1	-1	-2	-1	-1

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The tests of overall interaction are minimally affected by the difference in sample sizes portrayed in the two cases, except for Patterns 8 and 2, which have quantitative interaction. The differences in the other patterns are all less than or equal to 1%. In the patterns with no interaction or substantial qualitative interaction

(Patterns 1, 11, 5, and 10), all of the methods would give similar results if used as a preliminary test in a two-stage testing procedure. For the patterns with interaction, but little or no qualitative interaction (Patterns 8, 2 and 3), the sample size shift between the two centers has an important effect. For Pattern 4, only the test of Gail and Simon and the approximate method of Azzalini and Cox are substantially affected.

6.9 Discussion

6.9.1 Test of Overall Interaction and Methods for the Detection of Qualitative Interaction

It is clear from the results presented in this chapter that the reliability of the tests of qualitative interaction to accurately detect qualitative interaction depends both on the degree of interaction present in the data and the relative sample sizes at the two centers. For the cases presented in this study, the minimum and maximum percentages of simulations with significant tests of overall interaction and significant tests or substantial evidence of qualitative interaction are presented in Table 44.

A comparison of the results of Pattern 1 across the five sample size configurations provides an assessment of the methods when there is a significant treatment difference and no significant interaction. The error rate for the two tests of overall interaction is near the expected rate of 10% for all sample size configurations. The error rates for the tests of qualitative interaction range from 0% to 5%. The number of datasets with both positive and negative treatment differences in the raw data ranges from 4% to 15%.

When there is notable quantitative interaction, such as in Pattern 2, the tests of overall interaction find significant interaction in 88% of the simulations, when the sample size is balanced (Table 13). As the degree of imbalance increases, the number

Table 44

**Minimum and Maximum Percentages of Simulations With Significant Tests
or Substantial Evidence of Overall Interaction or Qualitative Interaction
for Two-Center Simulations**

Pattern and Int. Type*	Overall Interaction		Qualitative Interaction				
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	8, 11	8, 11	1, 4	1, 5	0, 1	0, 3	4, 15
8 QN	25, 39	25, 39	2, 10	1, 13	0, 5	0, 10	12, 28
2 QN	65, 88	66, 88	21, 23	9, 28	8, 11	5, 47	49, 50
3 QL	91, 99	92, 99	44, 70	48, 60	24, 47	26, 88	70, 90
4 QL	99, 100	99, 100	64, 97	70, 93	44, 89	56, 99	86, 100
5 QL	100, 100	100, 100	92, 100	95, 100	82, 100	89, 100	99, 100
11 NI	8, 11	8, 11	9, 11	4, 11	2, 2	1, 17	48, 51
10 QL	99, 100	99, 100	92, 100	94, 100	82, 99	89, 98	98, 100

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

of simulations with significant overall interaction decreases. In Cases 2.4 and 2.5, the number of significant interactions, based on these tests, is near 65% (Table 31 and Table 36). For this pattern, the effect of sample size imbalance on the methods of detecting of qualitative interaction differs markedly. Since this pattern does not display qualitative interaction, the numbers of significant tests of qualitative interaction detected by these tests can be considered as an error rate. The error rate for the test of Gail and Simon is near the expected rate of 10% across all sample size patterns. The exact method of Azzalini has a higher error rate but is also consistent,

ranging from 21% to 23%. The approximate method of Azzalini and Cox varies from a low of 9% to a high of 28%. The method of Ciminera et al. is by far the most variable. The error rate is 5% for the case of equal sample sizes (Table 14) and 47% for Case 2.2 (Table 19).

All of the methods of detecting qualitative interaction display substantial variability among sample size configurations for the patterns with qualitative interaction. The degree of variability is less for patterns such as Patterns 5 and 10, than for patterns with lesser degrees of qualitative interaction, such as Patterns 3 and 4. For Pattern 3, the number of simulations considered to show “substantial evidence” of qualitative interaction by the method of Ciminera et al. ranges from 26% in the equal case (Case 2.1, Table 14) to 30% in Case 2.3 (Table 24) to 88% in Cases 2.2 and 2.4 (Table 19 and Table 32). For Pattern 4, the range of the test of Gail and Simon is from 44% in Case 2.5 (Table 37) to 89% in Case 2.4 (Table 32). For this same pattern, the approximate method of Azzalini and Cox has a range of 70% in Case 2.5 to 90% in Case 2.4. The range in Pattern 4 for the exact test of Azzalini and Cox is from 64% in Case 2.5 to 97% in Case 2.4. For Ciminera et al., the range is from 56% in Case 2.5 to 99% in Case 2.4. This range is an indication of the sensitivity of the methods to the degree of qualitative interaction present in the data. Pattern 4 has a treatment effect of -0.5 at Center 1 and 1.5 at Center 2. For Case 2.4, with a sample size ratio of 54:10 for Center 1:Center 2, the degree of qualitative interaction is much higher than for Case 2.5 with a sample size ratio of 10:54.

The amount of variability in the results for the tests of overall interaction is much less for these patterns. The number of significant interactions for Pattern 4 is always greater than 99%. The equal sample size configuration of Pattern 3 (Table 13), as well as Cases 2.2 and 2.3 (Table 18 and Table 23), are always greater

than 99% and Cases 2.4 and 2.5 are always greater than 91% (Table 31 and Table 36).

Although there are some exceptions, the results for the two patterns without treatment differences are consistent across methods. Both methods of testing for overall interaction detected interaction in over 99% of the simulations for Pattern 10 across all sample size configurations. With three exceptions, the percentages of detected interactions were greater than 90% for all methods of detecting qualitative interaction across the sample sizes. Error rates for Pattern 11 were from 9% to 11% in the tests for overall interaction and 11% or below for the methods of detecting qualitative interaction, except for the method of Ciminera et al.

6.9.2 Test of Non-inferiority

The minimum and maximum percentages of the significant tests of non-inferiority, tested using the test of non-inferiority proposed by Gail and Simon, are presented in Table 45.

The error rate of the test is near the appropriate error level, 2.5%, for all sample size cases for Pattern 11. In this pattern the mean of each treatment group is equal to the other at each center. For Pattern 10, there is no overall treatment effect and each treatment is numerically, if not statistically, superior to the other at one of the two centers. The results for Pattern 10 reflect the relative sample sizes at Center 2, where the mean of Treatment 2 is less than the mean of Treatment 1 and at Center 1, where the mean of Treatment 2 is greater than the mean of Treatment 1. The percentages range from 46% or 47% to 100% for both treatments.

Table 45

**Minimum and Maximum Percentages of Simulations With a Significant Test That
One Treatment Group Mean is Not \geq the Other Treatment Group Mean
at Both Centers**

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at Both Centers	Treatment 1 is not \geq Treatment 2 at Both Centers
1	0, 0	73, 75
8	0, 0	47, 95
2	1, 2	48, 100
3	5, 14	66, 100
4	12, 60	83, 100
5	46, 100	98, 100
11	3, 4	2, 3
10	47, 100	46, 100

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

Pattern 11 is the only pattern where, for at least one of the centers, Treatment 2 is not numerically superior to Treatment 1 by at least 0.5, which represents the clinically relevant difference the study is powered to detect. For all patterns other than Patterns 10 and 11, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 is at least 69% for all cases except Case 2.4. For this case, the power for this test varies from 47% to 98% for these patterns. For Patterns 1, 8 and 2, the overall treatment effect is equal to the clinically relevant difference (with Treatment 2 being superior) and Treatment 2 is equal to or greater than Treatment 1 at both centers, i.e. there is either no interaction or quantitative interaction. For these

patterns, the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at both centers is less than or equal to 2%. For Pattern 3, where there is a qualitative interaction with the negative treatment effect, at Center 1, of one-half of the clinically relevant difference, the rejection rate of the test that Treatment 2 is equal to or greater than Treatment 1 at both centers varies from 5% to 14%. For Patterns 4 and 5, with stronger qualitative interaction, the power of this test varies greatly across sample size cases, ranging from 12% to 100%

6.9.3 Two-stage Test Results

The ranges of significant test results for the test of difference between the two treatment groups, when a two-stage testing procedure is used to select the final analysis model, are presented in Table 46.

A comparison of the results of Pattern 1 across the five sample size configurations provides an assessment of the methods when there is a significant treatment difference and no significant interaction. For Pattern 1, the power to detect the treatment difference through a two-stage testing procedure, using any of the tests of overall interaction or qualitative interaction as a pretest, is always as powerful as using a Type III model and has almost as much power as a Type II model. In the cases with greatest degree of imbalance, the power with a Type III model is only slightly better than an unbiased coin toss, whereas using a two-stage procedure provides power greater than 76%.

In the discussion of the case of equal sample sizes, we also noted that for patterns with interaction the power was near 80% using any of the methods as a pretest (Table 17). For the patterns with no interaction, the error rate was at the expected rate of 5% for all methods.

Table 46

Minimum and Maximum Percentages of Simulations With Significant Tests
That Treatment 2 is Not Equal to Treatment 1 After Model Selection
With a Preliminary Test of Interaction

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	76, 80	76, 80	78, 80	77, 80	79, 80	78, 80	71, 80
8 QN	55, 80	55, 80	44, 87	44, 86	44, 92	46, 89	45, 80
2 QN	49, 80	49, 80	19, 82	14, 80	14, 89	32, 90	31, 79
3 QL	52, 80	52, 80	34, 79	21, 79	23, 80	47, 80	46, 79
4 QL	52, 80	52, 80	50, 80	45, 80	46, 79	52, 79	52, 80
5 QL	52, 80	52, 80	52, 80	52, 80	52, 80	52, 80	52, 80
11 NI	5, 8	5, 8	4, 5	4, 5	4, 5	5, 6	5, 6
10 QL	4, 7	4, 7	5, 11	5, 9	5, 20	5, 13	5, 6

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

A more general evaluation of the two-stage testing procedure incorporating the methods of detecting qualitative interaction is not appropriate for the other cases of two-center simulations. The simulations presented in this study do not provide an accurate assessment of the method. The Type II treatment means are highly influenced by the respective treatment pattern and sample size. The power of the Type II analysis to detect between-treatment differences is a function of these treatment differences. And this power contributes to the power of the two-stage process through the selection of the Type II model by the first-stage test.

For example, in Case 2.4 Pattern 3, the Type II means are 0.000 and -0.018 for Treatments 1 and 2, respectively and the Type III means are 0.000 and 0.491 for Treatments 1 and 2, respectively (Table 34). The power of the test of differences between the two means is 4% for Type II and 52% for Type III. A two-stage procedure using the Gail and Simon procedure, which identifies significant qualitative interaction in 47% of the simulations (Table 32) has a power to detect significant treatment differences of 23% (Table 35). And in Case 2.5 Pattern 3, the Type II means are 0.000 and 1.013 for Treatments 1 and 2, respectively and the Type III means are 0.001 and 0.499 for Treatments 1 and 2, respectively (Table 39). The power of the test of differences between the two means is 100% for Type II and 53% for Type III. A two-stage procedure using the Gail and Simon procedure, which identifies significant qualitative interaction in 24% of the simulations (Table 37) has a power to detect significant treatment differences of 77% (Table 40).

CHAPTER VII

RESULTS FROM SIMULATED DATA FOR THREE CENTERS

7.1 Introduction

The operating characteristics of the methods of identifying quantitative interaction proposed by Azzalini and Cox (1984), Gail and Simon (1985) and Ciminera et al. (1993) were examined using simulated datasets depicting a three-center trial with two treatments.

As described in Chapter V, data were simulated for 11 patterns of treatment-by-center interaction and for seven sample size configurations for each interaction pattern. The total sample size for each treatment group was 64, but the distribution of the patients across the three centers differed for each configuration. The sample sizes of the two treatment groups within each center were always equal to each other. For each pattern and sample size configuration, 2500 datasets were simulated and summarized.

The 2500 datasets for each pattern and sample size configuration were generated using the same initial seed number and adjusting each generated number by an appropriate amount to produce the expected treatment-by-center means. Thus differences between the patterns and configurations are not the result of any differences in the sets of random numbers.

7.2 Case 3.1: 22 Patients at Center 1, 21 Patients at Center 2 and 21 Patients at Center 3

7.2.1 General Results

The simulations for the “equal” sample size configuration generated 22 observations for each treatment group at Center 1 and 21 for each treatment group at Centers 2 and 3. The average means and variances of these 2500 datasets, without any adjustment to the means for treatment and center effects, are presented in Table 47. The target value of each of these means was 0.000 and the target variance of the observations was 1.000.

Table 47

Means and Variances by Treatment and Center of Three-Center Simulated Data

Center	Treatment	Mean	Variance
1	1	0.0014	0.945
1	2	-0.0009	0.961
2	1	-0.0003	0.991
2	2	0.0009	0.990
3	1	-0.0002	1.004
3	2	-0.0086	0.987

The randomly generated numbers were consistent with the targeted values for the simulation.

7.2.2 Tests of Overall Interaction

The results of the test of significance of the treatment-by-center interaction from analysis of variance and using the H statistic proposed by Gail and Simon were obtained for each simulated dataset. The analysis of variance test was considered to be significant if the F statistic for the test had a significance level less than or equal to 10%. The H test was considered significant if the level of the associated chi-square test was less than or equal to 10%.

The percentage of significant results for each test from the 2500 simulated datasets of each interaction pattern is presented in Table 48. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

The two tests produce results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in 99% or more of the individual simulations for each pattern.

The power of the tests to detect interaction is at or near 100% for Patterns 5, 6, 7, 8 and 9. Patterns 6, 7, 8 and 9 are patterns that have an overall treatment effect and qualitative interaction. In Pattern 5, two of the three centers have no treatment effect and the third center has a treatment effect that is three times the clinically relevant difference that the trial is powered to detect. Significant interaction is detected over 65% of the time for Patterns 2 and 4, which have an overall treatment effect and quantitative interaction. The power of these tests to detect the weaker quantitative interaction (Pattern 3) is 53%. In the patterns with no interaction (Patterns 1 and 11), the error rate for the detection of interaction is near the expected error rate of 10%. The power to detect significant interaction in Pattern 10, which has no overall treatment difference, but has clinically relevant differences with different signs at two

of the centers, is lower (65% or 66%) than that for the other patterns with quantitative interaction.

Table 48

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.1

Pattern and Int. Type*	Characteristics of Simulation Patterns					Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Center 3 Effect	Analysis of Variance	G. and S. H Statistic
1 NI	Yes	None	0.5	0.5	0.5	8	9
2 QN	Yes	Quant.	0.2	0.2	1.1	66	66
3 QN	Yes	Quant.	0.0	0.75	0.75	53	54
4 QN	Yes	Quant.	0.0	0.5	1.0	65	66
5 QN	Yes	Quant.	0.0	0.0	1.5	98	98
6 QL	Yes	Qual.	-0.5	1.0	1.0	98	98
7 QL	Yes	Qual.	-0.5	0.5	1.5	99	100
8 QL	Yes	Qual.	-0.5	0.0	2.0	100	100
9 QL	Yes	Qual.	-0.25	-0.25	2.0	100	100
11 NI	No	None	0.0	0.0	0.0	8	9
10 QL	No	Qual.	-0.5	0.0	0.5	65	66

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

7.2.3 Methods for the Detection of Qualitative Interaction

The number of simulated datasets with qualitative interaction was evaluated using the tests of Azzalini and Cox and Gail and Simon, and the method of Ciminera et al. Both of the extended methods of Azzalini and Cox (exact and approximate methods) presented in Chapter IV were used. The methods of Gail and Simon and Ciminera et al. were evaluated using the variance estimate from the analysis of variance, as presented in Chapters III and IV.

The significance of the Azzalini and Cox test was calculated for each simulated dataset for both the exact and approximate procedures. Each dataset was tested for qualitative interaction using Gail and Simon's test with the minimum of Q^+ and Q^- . For each of these procedures, a dataset was considered to have a significant qualitative interaction if the level of the test was less than or equal to 10%. The number of significant tests from the 2500 simulated datasets was summarized.

The results of the method of Ciminera et al. were examined for each dataset. Any dataset with "pushed back" (destandardized) mean differences which were not all either positive or negative was considered to show "substantial evidence" of the presence of qualitative interaction. The number of datasets showing "substantial evidence" was summarized.

The results of the three methods are presented in Table 49, along with the percentages of datasets in which the treatment differences in the raw data have differing signs across the three centers. A frequently used, ad-hoc, method of evaluating data from multicenter trials for the presence of qualitative interaction is to examine the data for differing signs among the treatment differences. If all of the signs of the treatment differences are not either positive or negative, then the dataset is considered to have "substantial evidence" of qualitative interaction.

Since these methods are designed to detect qualitative interaction, the percentages of qualitative interactions detected in Patterns 1, 2, 3, 4, 5 and 11 will be considered to be error rates for these tests. The percentages of qualitative interactions detected in Patterns 6, 7, 8, 9 and 10 will be considered to be the power of the tests.

Table 49

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.1

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	1	1	0	0	14
2 QN	6	6	2	8	44
3 QN	13	12	4	3	50
4 QN	13	13	4	5	52
5 QN	24	24	10	27	75
6 QL	71	70	47	40	96
7 QL	71	70	48	43	96
8 QL	75	74	54	69	98
9 QL	62	61	39	70	96
11 NI	8	8	1	4	75
10 QL	55	55	28	39	95

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

As expected the exact and approximate methods of Azzalini and Cox provide equivalent results when sample sizes are equal. In the discussion of the results in this section, the two methods will not be differentiated.

The three methods have very different sensitivities to different patterns of interaction. For the two patterns where the means for Treatment 2 are greater than those for Treatment 1 at all centers (Patterns 1 and 2), the percentages of detected qualitative interactions are below the expected level of 10% for all methods. Patterns 3 and 4 have means for Treatment 2 that are greater than those for Treatment 1 at two centers and means that are equal for the two treatments at the other center. For these two patterns, the percentages of detected qualitative interactions are just above the expected level of 10% for Azzalini and Cox, but are much lower for the other two methods. Pattern 5 has two centers with means that are equal for both treatments and a third center where the means for Treatment 2 are greater than those for Treatment 1. The methods of Azzalini and Cox and Ciminera et al. each detect qualitative interaction in about 25% of the simulations, while the test of Gail and Simon has a level at the expected error rate of 10%.

For Patterns 6 and 7, the means for Treatment 2 are greater than those for Treatment 1 at two centers and the means for Treatment 2 are less than those for Treatment 1 at the other center. At this latter center, the difference is equal to the clinically significant difference. For these two patterns, the percentages of detected qualitative interactions are 70% for Azzalini and Cox, near 50% for Gail and Simon and slightly less for Ciminera et al. Pattern 8 has a large positive difference at one center, no difference at a second and a clinically significant negative difference at the third. The percentages of qualitative interactions detected range from 54% to 75%.

The range for Pattern 9, which has a large positive difference at one center and two smaller negative differences at the other two, is from 39% to 70%.

For Pattern 11, the treatment means are equal across treatments and centers. The error rates for Gail and Simon and Ciminera et al. are 1% and 4%, respectively, and 8% for Azzalini and Cox. Pattern 10 has equal treatment means at one center and treatment differences at the other two centers that are of equal magnitude, but opposite in sign. The numbers of detected qualitative interactions ranged from 28% to 55%.

All three methods provide a more realistic approach to determining the presence of qualitative interaction than examining the presence of differing signs in the raw data. In the patterns where there are positive treatment effects at two of the three centers and no treatment effect at the other (Patterns 3 and 4), the raw data show evidence of a qualitative interaction in approximately 50% of the simulations. This coincides with the expectation that 50% of the treatment effect values at the no effect center will be less than the mean of zero and 50% greater than zero. The detection level is also near 50% for Pattern 2, but is much larger for all other patterns, except Pattern 1.

7.2.4 Test of Non-inferiority

The null hypotheses that each respective treatment group is at least as good as the other treatment group at every center were tested using the test of non-inferiority proposed by Gail and Simon. The method was evaluated using the variance estimate from the analysis of variance, as presented in Chapters III and IV.

Each treatment group in each dataset was tested for non-inferiority against the other using Gail and Simon's test with both Q^+ and Q^- . A dataset was considered to

reject the null hypothesis that the mean of Treatment 1 is equal to or greater than the mean of Treatment 2 (or vice versa) at all centers if the level of the test was less than or equal to 2.5%. The number of significant tests from the 2500 simulated datasets was summarized.

The results of the methods are presented in Table 50.

In the pattern where the mean of each treatment group is equal to the other at each center (Pattern 11), the error rate of each test is near the appropriate error level of 2.5%. Pattern 11 is the only pattern where, for at least one of the centers, Treatment 2 is not numerically superior to Treatment 1 by at least 0.5, which represents the clinically relevant difference the study is powered to detect. For Pattern 10, there is no overall treatment effect and each treatment is numerically superior to the other at one of the three centers. The power of the test to show that a given treatment is not equal to or superior to the other at all three centers is 21% for Treatment 1 or 19% for Treatment 2, respectively.

For all patterns other than Patterns 11 and 10, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 at all centers is at least 71%. For Patterns 1, 2, 3, 4, and 5, the overall treatment effect is equal to the clinically relevant difference (with Treatment 2 being superior) and Treatment 2 is equal to or greater than Treatment 1 at all three centers, i.e. there is either no interaction or quantitative interaction. For these patterns, the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at all three centers is less than or equal to 1%. However, for the patterns with qualitative interaction (Patterns 6, 7, 8, 9 and 10), the power to detect that Treatment 2 is not \geq Treatment 1 at all three centers never exceeds 21%.

Table 50

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.1

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at All Centers	Treatment 1 is not \geq Treatment 2 at All Centers
1 NI	0	71
2 QN	0	90
3 QN	0	85
4 QN	0	88
5 QN	1	99
6 QL	17	99
7 QL	17	100
8 QL	21	100
9 QL	12	100
11 NI	3	2
10 QL	21	19

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

7.2.5 Two-stage Test Results

The purpose of two-stage testing is to improve the power and lower the error rate for the test of the overall treatment difference by using a preliminary test to select the appropriate final analysis model. The Type II model is selected as the final model when there is no consequential evidence that the interaction term needs to be included

in the model. The existence of such evidence justifies the inclusion of the interaction term in the model and the Type III model is selected as the final analysis model.

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 51.

For all patterns, the Type III and Type II overall treatment means for both treatments are near the targeted values of the simulation. This is the expected result for the case of equal sample sizes. There are small differences in the means from the two models for some patterns due to the slight imbalance of the sample sizes of the three centers. The smaller than expected Type II means for these patterns reflect the relatively larger sample size at Center 1, where Treatment 2 means are less than the respective Treatment 2 means at one or both of the other centers.

The tests of significance of the difference between the two means using the Type III model are consistent with the selected power and error rate. The power to detect differences, when they are present (all patterns except Patterns 11 and 10), is near 80%; and the error rate, when there is no between-treatment difference (Patterns 11 and 10), is at 5%. The results of the Type II model are similar to those of the Type III model when there is no interaction or minimal quantitative interaction (Patterns 1, 2, 3 and 4); however, the power decreases in correspondence with the degree of interaction. The power for Patterns 8 and 9, the most extreme patterns of qualitative interaction, is near 70%.

Table 52 summarizes test results when a two-stage testing procedure is used. The final analysis model contains the interaction term (Type III model) only if the preliminary test indicates that there is significant interaction (Analysis of Variance, Gail and Simon H statistic), significant qualitative interaction (Azzalini and Cox [both methods], Gail and Simon) or substantial evidence of qualitative interaction

(Ciminera et. al, Raw Data). Otherwise, the Type II model is the final analysis model.

Table 51

Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 3.1

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III	Type II	Type III	Type II	Type III	Type II
	Model	Model				
1 NI	80	80	0.000	0.000	0.497	0.497
2 QN	80	79	0.000	0.000	0.497	0.493
3 QN	80	78	0.000	0.000	0.497	0.489
4 QN	80	78	0.000	0.000	0.497	0.489
5 QN	80	75	0.000	0.000	0.497	0.489
6 QL	80	74	0.000	0.000	0.497	0.482
7 QL	80	73	0.000	0.000	0.497	0.482
8 QL	80	69	0.000	0.000	0.497	0.482
9 QL	80	70	0.000	0.000	0.497	0.485
11 NI	5	5	0.000	0.000	-0.003	-0.003
10 QL	5	5	0.000	0.000	-0.003	-0.011

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

Table 52

**Percentage of Simulations With Significant Test That Treatment 2
is Not Equal to Treatment 1 After Model Selection With
a Preliminary Test of Interaction for Case 3.1**

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	81	81	80	80	80	80	81
2 QN	80	80	79	79	79	79	80
3 QN	80	80	79	79	79	79	80
4 QN	80	80	79	79	79	79	80
5 QN	80	80	78	78	77	79	80
6 QL	80	80	80	80	79	79	80
7 QL	80	80	80	80	79	78	80
8 QL	80	80	80	80	79	80	80
9 QL	80	80	79	79	78	80	80
11 NI	5	5	5	5	5	5	5
10 QL	5	5	5	5	5	5	5

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The percentages of differences found to be significant do not differ substantially depending on the method used as a preliminary test. For those patterns with a significant treatment effect (all patterns except Patterns 11 and 10), the power

to detect the difference ranges from 78% to 81%. For Patterns 11 and 10, the error rate is 5% for all methods.

These results integrate the results for the case with equal sample sizes, presented in this section. The power of the Type III model to detect the treatment effect, when it is present, is always near 80%. The power of the Type II model is also near 80% when there is little evidence of interaction, but decreases as the amount of interaction increases. The power of the methods to detect interaction varies, but is generally higher in those patterns (Patterns 8 and 9) where there is a treatment effect and substantial qualitative interaction. In those patterns, the more powerful Type III model is chosen.

The error rates for Patterns 11 and 10, patterns without a treatment effect, are close to the expected 5% error rate.

7.3 Case 3.2: 39 Patients at Center 1, 13 Patients at Center 2 and 12 Patients at Center 3

The simulations for this unequal sample size configuration generated 39 observations for each treatment group at Center 1, 13 observations for each treatment group at Center 2 and 12 observations for each treatment group at Center 3. For this case, the evidence of qualitative interaction should be strong for Patterns 6, 7 and 8, for which Center 1 has a clinically significant negative treatment effect. For Pattern 9, Centers 1 and 2 both have weak negative effects; hence the evidence is less compelling.

7.3.1 Tests of Overall Interaction

The percentage of significant results for each test from the 2500 simulated datasets of each interaction pattern is presented in Table 53. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

Table 53

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.2

Pattern and Int. Type*	Characteristics of Simulation Patterns					Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Center 3 Effect	Analysis of Variance	G. and S. H Statistic
1 NI	Yes	None	0.5	0.5	0.5	10	11
2 QN	Yes	Quant.	0.2	0.2	1.1	52	53
3 QN	Yes	Quant.	0.0	0.75	0.75	55	56
4 QN	Yes	Quant.	0.0	0.5	1.0	61	61
5 QN	Yes	Quant.	0.0	0.0	1.5	90	91
6 QL	Yes	Qual.	-0.5	1.0	1.0	98	98
7 QL	Yes	Qual.	-0.5	0.5	1.5	99	99
8 QL	Yes	Qual.	-0.5	0.0	2.0	100	100
9 QL	Yes	Qual.	-0.25	-0.25	2.0	100	100
11 NI	No	None	0.0	0.0	0.0	10	11
10 QL	No	Qual.	-0.5	0.0	0.5	61	61

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The two tests provide results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in 99% or more of the individual simulations for each pattern.

The power of the tests to detect interaction in the patterns with qualitative interaction and an overall treatment effect (Patterns 6, 7, 8 and 9) is at or near 100%. Pattern 10 has qualitative interaction and no overall treatment effect, interaction was detected in 61% of the simulations. The power to detect interaction in patterns with quantitative interaction (Pattern 2, 3, 4 and 5) ranges from 52% to 91%. In the patterns with no interaction (Patterns 1 and 11), the error rate for both methods is near the expected error rate of 10%.

7.3.2 Methods for the Detection of Qualitative Interaction

The percentages of simulated datasets with qualitative interaction, as evaluated using the tests of Azzalini and Cox (exact and approximate methods) and Gail and Simon, and the method of Ciminera et al., are presented in Table 54, along with the percentages of datasets with differing signs in the raw data.

The results from the exact and approximate approaches of Azzalini and Cox are different in this case. The exact approach generally has a higher rejection rate than the approximate approach.

The three methods provide similar results in the pattern with no interaction and an overall treatment difference (Pattern 1). For the other patterns with overall treatment differences (all other patterns except Patterns 10 and 11), the number of detected qualitative interactions is consistently higher for the method of Ciminera et al., consistently lower for the test of Gail and Simon and the test of Azzalini and Cox

is between the other two. The presence of differing signs in the raw data generally has a higher rejection rate than any of the three proposed methods.

Table 54

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.2

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	2	2	0	2	22
2 QN	7	7	2	19	43
3 QN	12	4	3	47	53
4 QN	13	5	3	47	55
5 QN	24	19	10	53	76
6 QL	86	69	68	98	99
7 QL	86	70	69	99	99
8 QL	88	75	74	99	99
9 QL	65	52	44	90	97
11 NI	9	8	1	18	75
10 QL	52	48	27	50	93

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For three of the patterns with an overall treatment effect and qualitative interaction (Patterns 6, 7, 8), the positive indications of qualitative interaction are

high for the method of Ciminera et al. (98% to 99%) and the exact method of Azzalini and Cox (86% to 88%). The rates for the approximate method of Azzalini and Cox (69% to 75%) and the test of Gail and Simon (68% to 74%) are lower. For the other pattern with an overall treatment effect and qualitative interaction (Pattern 9), the number of significant tests is lower for all tests, ranging from 44% to 90%.

For Pattern 2, which has one large difference and two smaller differences, all with the same sign, the error rates for the tests of Gail and Simon (2%) and Azzalini and Cox (7%, 7%) are less than the expected level (10%). For this pattern, the method of Ciminera et al. finds “substantial evidence” of qualitative interaction 19% of the time. Patterns 3 and 4 each have one center with no between-treatment difference and two centers with differences with the same sign. The error rates for the tests of Gail and Simon (3%, 3%) and the approximate method of Azzalini and Cox (4%, 5%) are less than the expected level (10%). The level of the exact method of Azzalini and Cox (12%, 13%) is higher, but still much less than that of the method of Ciminera et al. (47%, 47%). The error rates for Pattern 5, which has no between-treatment differences at two centers and a difference at the third, range from the expected level of 10% for the test of Gail and Simon to 53% for the method of Ciminera et al.

In the pattern with no interaction, and no treatment effect (Pattern 11), the error rates of the Azzalini and Cox are near the expected level of 10%, the level of Ciminera et al. is highest (18%), and that of Gail and Simon test is much less (1%). For Pattern 10, the detection rate for Gail and Simon is 27%, and that for the other methods is near 50%.

7.3.3 Test of Non-inferiority

The percentages of significant tests of the null hypotheses that each respective treatment group is at least as good as the other treatment group at every center, tested using the test of non-inferiority proposed by Gail and Simon, are presented in Table 55.

In the pattern where the mean of each treatment group is equal to the other at each center (Pattern 11), the error rate of the test is near the appropriate error level of 2.5%. Pattern 11 is the only pattern where, for at least one of the centers, Treatment 2 is not numerically superior to Treatment 1 by at least 0.5, which represents the clinically relevant difference the study is powered to detect. For Pattern 10, there is no overall treatment effect and each treatment is numerically superior to the other at one of the three centers. The power of the test to show that a given treatment is not equal to or superior to the other at all three centers is 11% for Treatment 1 and 38% for Treatment 2. The difference in these two results is due to the unequal sample size at the two centers where the treatment means are equal in magnitude but differ in signs.

For all patterns other than Patterns 11 and 10, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 is at least 61%. For Patterns 1, 2, 3, 4, and 5, the overall treatment effect is equal to the clinically relevant difference (with Treatment 2 being superior) and Treatment 2 is equal to or greater than Treatment 1 at all three centers, i.e. there is either no interaction or there is quantitative interaction. For these patterns, the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at all three centers is less than or equal to 1%. However, for the patterns with qualitative interaction (Patterns 6, 7, 8, 9 and 10),

the power to detect that Treatment 2 is not \geq Treatment 1 at all three centers never exceeds 37%.

Table 55

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.2

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at All Centers	Treatment 1 is not \geq Treatment 2 at All Centers
1 NI	0	71
2 QN	0	69
3 QN	0	61
4 QN	0	65
5 QN	1	88
6 QL	33	85
7 QL	33	91
8 QL	37	99
9 QL	14	99
11 NI	3	2
10 QL	38	11

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

7.3.4 Two-stage Test Results

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 56.

Table 56

Treatment Means for Each Treatment Calculated From Type II and Type III
Analyses and the Percentages of Simulations With Significant Tests
That Treatment 2 is Not Equal to Treatment 1 for Case 3.2

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III	Type II	Type III	Type II	Type III	Type II
	Model	Model				
1 NI	68	80	0.001	0.000	0.494	0.497
2 QN	68	52	0.001	0.000	0.494	0.366
3 QN	68	36	0.001	0.000	0.494	0.290
4 QN	68	35	0.001	0.000	0.494	0.286
5 QN	68	31	0.001	0.000	0.494	0.278
6 QL	68	5	0.001	0.000	0.494	0.083
7 QL	68	4	0.001	0.000	0.494	0.075
8 QL	68	4	0.001	0.000	0.494	0.067
9 QL	68	11	0.001	0.000	0.494	0.169
11 NI	5	5	0.001	0.000	-0.006	-0.003
10 QL	5	21	0.001	0.000	-0.006	-0.214

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

There are notable differences in the means from the two models for Treatment 2, only the means for Patterns 1 and 11 do not differ substantially. The smaller than expected Type II means for these patterns reflect the relatively larger sample size at

Center 1, where Treatment 2 means are less than the respective Treatment 2 means at one or both of the other centers.

The tests of significance of the difference between the two treatment means are consistent with the results of the calculation of the treatment means for the two models. The test results for the two models are only consistent for Pattern 11. For patterns with an overall treatment effect and some form of interaction (all except Patterns 1, 10, 11), the number of significant test results is higher for the Type III model than for the Type II model. For all of these patterns, the Type II means for Treatment 2 are less than the respective Type III means. For Pattern 10, with interaction but no treatment difference (in the Type III model), the number of significant tests is much higher for the Type II model than for the Type III model. The results of the test for this pattern reflect the larger sample size at Center 1, where the mean of Treatment 1 is less than the mean of Treatment 2.

Table 57 summarizes test results when a two-stage testing procedure is used to select the final analysis model.

For patterns with no interaction (Patterns 1 and 11), the percentages of differences found to be significant did not differ substantially depending on the method used as a preliminary test.

For those patterns with a significant treatment effect and interaction (all patterns except Patterns 1, 11, 10), the power to detect the difference was generally highest if one of the tests for overall interaction was used as a preliminary test. Using these preliminary tests, the power to detect treatment differences ranged from 60% to 68%. Of the tests for qualitative interaction, the method of Ciminera provided the best power, from 56% to 68%. The power for the tests of Azzalini and Cox and Gail and Simon ranged from 31% to 60% for these patterns. The patterns with the largest

differences between methods were Patterns 5 and 9. Both of these patterns have a large treatment difference at Center 3 and equal differences with no difference or a small opposite sign at Centers 1 and 2.

Table 57

Percentage of Simulations With Significant Test That Treatment 2
is Not Equal to Treatment 1 After Model Selection With
a Preliminary Test of Interaction for Case 3.2

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	79	78	80	80	80	80	74
2 QN	65	65	53	52	53	60	62
3 QN	60	61	41	37	37	59	59
4 QN	62	62	40	36	36	58	59
5 QN	67	67	39	35	33	56	63
6 QL	67	67	60	47	47	68	67
7 QL	68	68	59	46	47	67	67
8 QL	68	68	60	48	48	67	68
9 QL	68	68	47	36	31	64	67
11 NI	5	5	5	5	5	5	4
10 QL	11	10	16	14	19	19	7

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The power for Pattern 1 is at or near the expectation of 80% for all patterns. The results for Pattern 11 reflect the expected error rate of 5%. The error rate for Pattern 10 is higher (10% to 19%) due to the relatively larger sample size at Center 1.

7.4 Case 3.3: 13 Patients at Center 1, 39 Patients at Center 2 and 12 Patients at Center 3

The simulations for this unequal sample size configuration generated 13 observations for each treatment group at Center 1, 39 observations for each treatment group at Center 2 and 12 observations for each treatment group at Center 3. For this case, the evidence of qualitative interaction should not be strong for Patterns 6, 7 and 8, for which Center 2 has a positive or null treatment effect. For Pattern 9, Centers 1 and 2 both have weak negative effects; hence the evidence is stronger.

7.4.1 Tests of Overall Interaction

The percentage of significant results for each test from the 2500 simulated datasets of each interaction pattern is presented in Table 58. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

The two tests provide results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in 99% or more of the individual simulations for each pattern.

The power of the tests to detect interaction in the patterns with qualitative interaction and strong quantitative interaction when there is a significant overall treatment difference (Patterns 5, 6, 7, 8 and 9) is greater than 90%. The power to detect weaker patterns of quantitative interaction (Pattern 2, 3, 4) ranges from 42% to 53%. In the patterns with no interaction (Patterns 1 and 11), the error rate for both

methods is at the expected error rate of 10%. The power to detect the interaction when there is no overall significant treatment difference (Pattern 10) is less than 50%. This result reflects the smaller sample sizes at Centers 1 and 3.

Table 58

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.3

Pattern and Int. Type*	Characteristics of Simulation Patterns					Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Center 3 Effect	Analysis of Variance	G. and S. H Statistic
1 NI	Yes	None	0.5	0.5	0.5	10	10
2 QN	Yes	Quant.	0.2	0.2	1.1	52	53
3 QN	Yes	Quant.	0.0	0.75	0.75	42	43
4 QN	Yes	Quant.	0.0	0.5	1.0	44	45
5 QN	Yes	Quant.	0.0	0.0	1.5	90	91
6 QL	Yes	Qual.	-0.5	1.0	1.0	92	92
7 QL	Yes	Qual.	-0.5	0.5	1.5	93	94
8 QL	Yes	Qual.	-0.5	0.0	2.0	100	100
9 QL	Yes	Qual.	-0.25	-0.25	2.0	100	100
11 NI	No	None	0.0	0.0	0.0	10	10
10 QL	No	Qual.	-0.5	0.0	0.5	44	45

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

7.4.2 Methods for the Detection of Qualitative Interaction

The percentages of simulated datasets with qualitative interaction, as evaluated using the tests of Azzalini and Cox (exact and approximate methods) and Gail and Simon, and the method of Ciminera et al., are presented in Table 59, along with the percentages of datasets with differing signs in the raw data.

Table 59

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.3

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	2	3	0	2	22
2 QN	7	8	2	20	44
3 QN	13	16	5	8	52
4 QN	13	16	5	9	51
5 QN	24	19	10	53	74
6 QL	56	60	34	42	90
7 QL	56	60	34	43	90
8 QL	61	62	40	70	95
9 QL	65	53	44	90	96
11 NI	9	7	1	18	74
10 QL	37	38	15	43	90

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The results from the exact and approximate approaches of Azzalini and Cox are different in this case. The exact approach generally has a lower rejection rate than the approximate approach.

The three methods provide similar results in Pattern 1. For the other patterns with overall treatment differences (all other patterns except Patterns 10 and 11), the number of detected qualitative interactions is consistently lower for the test of Gail and Simon. The presence of differing signs in the raw data always has a higher rejection rate than any of the three proposed methods. For the methods of Azzalini and Cox and Ciminera et al., there is no consistency of which is higher than the other.

For two of the patterns with an overall treatment effect and qualitative interaction (Patterns 6 and 7), the positive indications of qualitative interaction are higher for the methods of Azzalini and Cox (56% to 60%) than for Ciminera et al. (42%, 43%). For the other patterns with an overall treatment effect and qualitative interaction (Patterns 8 and 9), the number of significant tests is higher for Ciminera et al. (70%, 80%) than for Azzalini and Cox (53% to 65%). The results of the test of Gail and Simon for these four patterns range from 34% to 44%.

For Pattern 2, the error rates for the tests of Gail and Simon (2%) and Azzalini and Cox (7%, 8%) are less than the expected level (10%). For this pattern, the method of Ciminera et al. finds “substantial evidence” of qualitative interaction 20% of the time. Patterns 3 and 4 each have error rates for the tests of Gail and Simon (5%, 5%) and the method of Ciminera et al. (8%, 9%) that are less than the expected level (10%). The levels of the methods of Azzalini and Cox (13%, 16%) are higher. The error rates for Pattern 5, range from the expected level of 10% for the test of Gail and Simon to 53% for the method of Ciminera et al.

In Pattern 11, the error rates of the Azzalini and Cox are near the expected level of 10%, the level of Ciminera et al. is highest (18%), and that of Gail and Simon test is much less (1%). For Pattern 10, the detection rate for Gail and Simon is 15%, and that for the other methods is near 40%.

7.4.3 Test of Non-inferiority

The percentages of significant tests of the null hypotheses that each respective treatment group is at least as good as the other treatment group at every center, tested using the test of non-inferiority proposed by Gail and Simon are presented in Table 60.

In Pattern 11, the mean of each treatment group is equal to the other at each center and the error rate of the test is near the appropriate error level of 2.5%. For Pattern 10, the power of the test to show that neither treatment is equal to or superior to the other at all three centers is near 10%.

For all patterns other than Patterns 11 and 10, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 is at least 71%. For Patterns 1, 2, 3, 4, and 5 (patterns with no interaction or quantitative interaction), the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at all three centers is less than or equal to 2%. However, for the patterns with qualitative interaction (Patterns 6, 7, 8, 9 and 10), the power to detect that Treatment 2 is not \geq Treatment 1 at all three centers never exceeds 14%.

7.4.4 Two-stage Test Results

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 61.

Table 60

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.3

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at All Centers	Treatment 1 is not \geq Treatment 2 at All Centers
1 NI	0	71
2 QN	0	71
3 QN	1	92
4 QN	1	82
5 QN	2	88
6 QL	10	99
7 QL	10	96
8 QL	12	99
9 QL	14	99
11 NI	3	2
10 QL	13	10

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

There are notable differences in the means from the two models for Treatment 2, although the means for several patterns (Patterns 1, 4, 7, 10 and 11) do not differ substantially. The differences in the Type II means for the other patterns reflect the relatively larger sample size at Center 2. For Patterns 3 and 6, the Treatment 2 means are greater than expected and for Patterns 2, 5, 8 and 9, Treatment 2 means are less than the expected.

Table 61

Treatment Means for Each Treatment Calculated From Type II and Type III
Analyses and the Percentages of Simulations With Significant Tests
That Treatment 2 is Not Equal to Treatment 1 for Case 3.3

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III Model	Type II Model	Type III	Type II	Type III	Type II
1 NI	66	80	0.000	0.000	0.496	0.498
2 QN	66	53	0.000	0.000	0.496	0.366
3 QN	66	92	0.000	0.000	0.496	0.595
4 QN	66	77	0.000	0.000	0.496	0.490
5 QN	66	31	0.000	0.000	0.496	0.279
6 QL	66	97	0.000	0.000	0.496	0.693
7 QL	66	73	0.000	0.000	0.496	0.482
8 QL	66	27	0.000	0.000	0.496	0.271
9 QL	66	12	0.000	0.000	0.496	0.170
11 NI	5	6	0.000	0.000	-0.004	-0.002
10 QL	5	5	0.000	0.000	-0.004	-0.010

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The tests of significance of the difference between the two treatment means are generally consistent with the results of the calculation of the treatment means for the two models. For patterns with larger than expected Type II means for Treatment 2 (Patterns 3 and 6), the Type II test is more powerful. For patterns with smaller than

expected Type II means (Patterns 2, 5, 8 and 9), the Type III test is more powerful. However, for patterns with an overall treatment effect and similar Type III and Type II means (Patterns 1, 4, 7), the Type II model is more powerful. For Patterns 11 and 10, the test results for the two models are consistent.

Table 62 summarizes test results when a two-stage testing procedure is used to select the final analysis model.

For patterns with no interaction and/or no treatment effect (Patterns 1, 11 and 10), the percentages of differences found to be significant did not differ substantially depending on the method used as a preliminary test. Similarly, for Patterns 4 and 7 the results were similar among the preliminary tests. The calculation of the Treatment 2 means for these two patterns is not greatly affected by the unequal sample size.

For those patterns with a significant treatment effect and interaction (all patterns except Patterns 1, 11, 10), the preliminary test with the most power to detect the difference was varied. For Patterns 2, 5, 8 and 9, the treatment effect at the center with the largest sample size (Center 2) was either zero or small in magnitude. For these patterns, the power was highest (65% to 66%) if one of the tests for overall interaction was used as a preliminary test. Of the tests for qualitative interaction, the method of Ciminera provided the best power, from 56% to 63%. The power for the tests of Azzalini and Cox and Gail and Simon ranged from 31% to 53% for these patterns. The pattern with the largest differences among methods was Pattern 5. For Patterns 3 and 6, patterns with the largest treatment difference at Center 2, the tests of qualitative interaction provided the best preliminary tests.

The power for Pattern 1 is at or near the expectation of 80% for all patterns. The results for Patterns 11 and 10 reflect the expected error rate of 5%.

Table 62

Percentage of Simulations With Significant Test That Treatment 2
is Not Equal to Treatment 1 After Model Selection With
a Preliminary Test of Interaction for Case 3.3

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	77	77	79	79	79	79	73
2 QN	65	65	53	53	53	61	61
3 QN	77	76	84	83	89	87	71
4 QN	72	72	73	73	76	75	68
5 QN	65	65	39	35	33	56	62
6 QL	67	67	71	70	78	75	66
7 QL	67	67	66	66	68	67	66
8 QL	66	66	50	49	39	59	66
9 QL	66	66	46	36	31	63	66
11 NI	6	6	6	5	6	6	5
10 QL	6	6	5	5	5	6	5

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

7.5 Case 3.4: 13 Patients at Center 1, 12 Patients at Center 2 and 39 Patients at Center 3

The simulations for this unequal sample size configuration generated 13 observations for each treatment group at Center 1, 12 observations for each treatment group at Center 2 and 39 observations for each treatment group at Center 3. For this

case, the evidence of qualitative interaction should not be strong for Patterns 6, 7, 8, and 9, for which Center 3 has a positive effect.

7.5.1 Tests of Overall Interaction

The percentage of significant results for each test from the 2500 simulated datasets of each interaction pattern is presented in Table 63. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

The two tests provide results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in 99% or more of the individual simulations for each pattern.

The power of the tests to detect interaction in the patterns with qualitative interaction (Patterns 6, 7, 8 and 9) is greater than 90%. The power to detect patterns of quantitative interaction (Pattern 2, 3, 4 and 5) ranges from 43% to 98%. In the patterns with no interaction (Patterns 1 and 11), the error rate for both methods is at or near the expected error rate of 10%.

7.5.2 Methods for the Detection of Qualitative Interaction

The percentages of simulated datasets with qualitative interaction, as evaluated using the tests of Azzalini and Cox (exact and approximate methods) and Gail and Simon, and the method of Ciminera et al., are presented in Table 64, along with the percentages of datasets with differing signs in the raw data.

The results from the exact and approximate approaches of Azzalini and Cox are different in this case. The exact approach generally has a lower rejection rate than the approximate approach.

Table 63

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.4

Pattern and Int. Type*	Characteristics of Simulation Patterns					Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Center 3 Effect	Analysis of Variance	G. and S. H Statistic
1 NI	Yes	None	0.5	0.5	0.5	10	11
2 QN	Yes	Quant.	0.2	0.2	1.1	70	71
3 QN	Yes	Quant.	0.0	0.75	0.75	43	44
4 QN	Yes	Quant.	0.0	0.5	1.0	63	63
5 QN	Yes	Quant.	0.0	0.0	1.5	98	98
6 QL	Yes	Qual.	-0.5	1.0	1.0	92	92
7 QL	Yes	Qual.	-0.5	0.5	1.5	99	100
8 QL	Yes	Qual.	-0.5	0.0	2.0	100	100
9 QL	Yes	Qual.	-0.25	-0.25	2.0	100	100
11 NI	No	None	0.0	0.0	0.0	10	11
10 QL	No	Qual.	-0.5	0.0	0.5	63	63

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The three methods provide similar results in Pattern 1. For the other patterns with overall treatment differences (all other patterns except Patterns 10 and 11), the number of detected simulations is consistently lowest for the test of Gail and Simon, consistently highest for the methods of Azzalini and Cox, with Ciminera et al. in the

middle. The presence of differing signs in the raw data always has a higher rejection rate than any of the three proposed methods.

Table 64

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.4

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	2	2	0	2	21
2 QN	10	13	3	6	53
3 QN	13	16	5	8	51
4 QN	14	17	5	9	55
5 QN	24	30	10	17	75
6 QL	56	60	34	42	90
7 QL	57	61	34	43	91
8 QL	61	67	40	52	95
9 QL	54	60	30	45	93
11 NI	9	7	2	18	76
10 QL	54	48	30	51	94

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For those patterns with an overall treatment effect and qualitative interaction (Patterns 6, 7, 8, and 9), the rates of detection of qualitative interaction ranged from

30% to 40% for the test of Gail and Simon to 42% to 52% for the method of Ciminera et al. to 56% to 67% for the methods of Azzalini and Cox.

The error rates for the patterns with quantitative interaction (Patterns 2, 3, 4 and 5) are all 10% or lower for the test of Gail and Simon. For the method of Ciminera et al., only Pattern 5 exceeds 10%, whereas for the Azzalini and Cox methods, only Pattern 2 is near 10%.

In Pattern 11, the error rates of the Azzalini and Cox are near the expected level of 10%, the level of Ciminera et al. is highest (18%), and that of Gail and Simon test is much less (2%). For Pattern 10, the detection rate for Gail and Simon is 30%, and that for the other methods is near 50%.

7.5.3 Test of Non-inferiority

The percentages of significant tests of the null hypotheses that each respective treatment group is at least as good as the other treatment group at every center, tested using the test of non-inferiority proposed by Gail and Simon are presented in Table 65.

In the pattern where the mean of each treatment group is equal to the other at each center (Pattern 11), the error rate of the test is near the appropriate error level of 2.5%. For Pattern 10, the power of the test to show that neither treatment is equal to or superior to the other at all three centers is 38% for Treatment 1 and 12% for Treatment 2. The difference in these two results is due to the unequal sample size at the two centers where the treatment means are equal in magnitude but differ in signs.

For all patterns other than Patterns 11 and 10, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 is at least 71%. For Patterns 1, 2, 3, 4, and 5 (patterns with no interaction or quantitative interaction), the error rate

of not finding Treatment 2 to be equal to or greater than Treatment 1 at all three centers is less than or equal to 2%. However, for the patterns with qualitative interaction (Patterns 6, 7, 8, 9 and 10), the power to detect that Treatment 2 is not \geq Treatment 1 at all three centers never exceeds 12%.

Table 65

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.4

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at All Centers	Treatment 1 is not \geq Treatment 2 at All Centers
1 NI	0	71
2 QN	0	99
3 QN	1	92
4 QN	1	98
5 QN	2	100
6 QL	10	99
7 QL	10	100
8 QL	12	100
9 QL	8	100
11 NI	3	3
10 QL	12	38

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

7.5.4 Two-stage Test Results

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 66.

Table 66

Treatment Means for Each Treatment Calculated From Type II and Type III Analyses and the Percentages of Simulations With Significant Tests That Treatment 2 is Not Equal to Treatment 1 for Case 3.4

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III	Type II	Type III	Type II	Type III	Type II
	Model	Model				
1 NI	69	80	0.000	0.000	0.499	0.497
2 QN	69	98	0.000	0.000	0.499	0.746
3 QN	69	91	0.000	0.000	0.499	0.595
4 QN	69	97	0.000	0.000	0.499	0.701
5 QN	69	100	0.000	0.000	0.499	0.912
6 QL	69	97	0.000	0.000	0.499	0.693
7 QL	69	100	0.000	0.000	0.499	0.904
8 QL	69	100	0.000	0.000	0.499	1.115
9 QL	69	100	0.000	0.000	0.499	1.119
11 NI	5	6	0.000	0.000	-0.001	-0.003
10 QL	5	19	0.000	0.000	-0.001	0.201

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

There are notable differences in the means from the two models for Treatment 2, only the means for Patterns 1 and 11 do not differ substantially. The larger than expected Type II means for these patterns reflect the relatively larger sample size at Center 3, where Treatment 2 means are larger than the respective Treatment 2 means at one or both of the other centers.

The tests of significance of the difference between the two treatment means are consistent with the results of the calculation of the treatment means for the two models. The test results for the two models are only consistent for Pattern 11. For all patterns except Pattern 11, the number of significant test results is higher for the Type II model than for the Type III model. For all of these patterns except Pattern 1, the Type II means for Treatment 2 are greater than the respective Type III means.

Table 67 summarizes test results when a two-stage testing procedure is used to select the final analysis model.

For patterns with no interaction (Patterns 1 and 11), the percentages of differences found to be significant did not differ substantially depending on the method used as a preliminary test.

For those patterns with a significant treatment effect and interaction (all patterns except Patterns 1, 11, 10), the power to detect the difference was generally highest if one of the methods for the detection of qualitative interaction was used as a preliminary test. Their power ranged from 69% to 96%.

Using the tests of overall interaction as preliminary tests, the power to detect treatment differences ranged from 69% to 78%. Of the tests for qualitative interaction, the test of Gail and Simon consistently provided the best power, from 78% to 96%. The power for the methods of Azzalini and Cox and Ciminera et al. ranged from 69% to 94% for these patterns. The differences among methods ranged

from 8% to 12%. The pattern with the largest difference among methods is Pattern 5, which has a large treatment difference at Center 3 and no difference at Centers 1 and 2.

Table 67

Percentage of Simulations With Significant Test That Treatment 2
is Not Equal to Treatment 1 After Model Selection With
a Preliminary Test of Interaction for Case 3.4

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	79	79	80	80	80	81	74
2 QN	72	72	91	89	96	94	71
3 QN	78	78	84	83	89	87	72
4 QN	74	74	88	86	94	91	71
5 QN	69	69	82	79	91	86	69
6 QL	70	70	73	72	80	76	69
7 QL	69	69	73	72	80	76	69
8 QL	69	69	70	69	75	71	69
9 QL	69	69	71	70	78	72	69
11 NI	6	7	6	6	6	7	6
10 QL	10	10	15	13	17	18	7

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The power for Pattern 1 is at or near the expectation of 80% for all patterns. The results for Pattern 11 reflect the expected error rate of 5%. The error rate for Pattern 10 is higher (10% to 18%) due to the relatively larger sample size at Center 3.

7.6 Case 3.5: 29 Patients at Center 1, 29 Patients at Center 2 and 6 Patients at Center 3

The simulations for this unequal sample size configuration generated 29 observations for each treatment group at Center 1, 29 observations for each treatment group at Center 2 and 6 observations for each treatment group at Center 3. For this case, the evidence of qualitative interaction should be strong for Patterns 6, 7 and 8, for which Center 1 has a negative treatment effect and Center 2 has a positive or null treatment effect. For Pattern 9, Centers 1 and 2 both have weak negative effects; hence the evidence is weaker.

7.6.1 Tests of Overall Interaction

The percentage of significant results for each test from the 2500 simulated datasets of each interaction pattern is presented in Table 68. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

The two tests provide results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in 99% or more of the individual simulations for each pattern.

The power of the tests to detect interaction in the patterns with qualitative interaction and strong quantitative interaction (Patterns 6, 7, 8 and 9) is equal to or greater than 95%. The power to detect patterns of quantitative interaction (Pattern 2,

3, 4 and 5) ranges from 37% to 72%. In the patterns with no interaction (Patterns 1 and 11), the error rate for both methods is near the expected error rate of 10%.

Table 68

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.5

Pattern and Int. Type*	Characteristics of Simulation Patterns					Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Center 3 Effect	Analysis of Variance	G. and S. H Statistic
1 NI	Yes	None	0.5	0.5	0.5	11	11
2 QN	Yes	Quant.	0.2	0.2	1.1	37	38
3 QN	Yes	Quant.	0.0	0.75	0.75	58	59
4 QN	Yes	Quant.	0.0	0.5	1.0	48	49
5 QN	Yes	Quant.	0.0	0.0	1.5	71	72
6 QL	Yes	Qual.	-0.5	1.0	1.0	99	99
7 QL	Yes	Qual.	-0.5	0.5	1.5	95	96
8 QL	Yes	Qual.	-0.5	0.0	2.0	97	97
9 QL	Yes	Qual.	-0.25	-0.25	2.0	96	96
11 NI	No	None	0.0	0.0	0.0	11	11
10 QL	No	Qual.	-0.5	0.0	0.5	48	49

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

7.6.2 Methods for the Detection of Qualitative Interaction

The percentages of simulated datasets with qualitative interaction, as evaluated using the tests of Azzalini and Cox (exact and approximate methods) and Gail and Simon, and the method of Ciminera et al., are presented in Table 69, along with the percentages of datasets with differing signs in the raw data.

Table 69

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.5

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	2	4	1	3	24
2 QN	6	1	1	5	40
3 QN	13	5	4	5	56
4 QN	12	4	3	4	54
5 QN	23	8	8	25	75
6 QL	80	61	59	44	98
7 QL	79	59	57	42	98
8 QL	81	61	62	68	99
9 QL	67	40	45	74	97
11 NI	9	4	2	9	76
10 QL	42	36	19	40	90

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The results from the exact and approximate approaches of Azzalini and Cox are different in this case. The exact approach generally has a higher rejection rate than the approximate approach.

The three methods provide similar results in the pattern with no interaction and an overall treatment difference (Pattern 1). However, test results differ markedly among the three for the other patterns. The presence of differing signs in the raw data consistently has a higher rejection rate than any of the three proposed methods.

For the patterns with quantitative interaction (Patterns 2, 3, 4 and 5), error rates were below 10% for the approximate method of Azzalini and Cox and the test of Gail and Simon. For the method of Ciminera et al., the error rate was 5% or below for all of these patterns except Pattern 5 (25%). The error rate for the exact method of Azzalini and Cox ranged from 6% to 23%. For Patterns 6, 7 and 8, the number of detected qualitative interactions is consistently lowest for the method of Ciminera et al. (44% to 68%), consistently highest for the exact method of Azzalini and Cox (79% to 81%) and intermediate for the test of Gail and Simon (57% to 62%) and the approximate method of Azzalini and Cox (59% to 61%). For Pattern 9, the number of detected qualitative interactions is highest for the method of Ciminera et al. and the exact method of Azzalini and Cox (74%, 67%). The test of Gail and Simon and the approximate method of Azzalini and Cox are lower (45%, 40%).

In the pattern with no interaction, and no treatment effect (Pattern 11), the error rates of the exact method of Azzalini and Cox and Ciminera et al. methods are highest (9%), but seem appropriate for a test with a level of 10%. The error rates of the approximate method of Azzalini and Cox and the Gail and Simon test are much less (4%, 2%). For Pattern 10, the detection rate for Gail and Simon is 19%, and that for the other methods is near 40%.

7.6.3 Test of Non-inferiority

The percentages of significant tests of the null hypotheses that each respective treatment group is at least as good as the other treatment group at every center, tested using the test of non-inferiority proposed by Gail and Simon, are presented in Table 70.

Table 70

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.5

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at All Centers	Treatment 1 is not \geq Treatment 2 at All Centers
1 NI	0	71
2 QN	0	49
3 QN	0	78
4 QN	0	62
5 QN	1	59
6 QL	24	95
7 QL	24	79
8 QL	29	82
9 QL	16	81
11 NI	2	3
10 QL	29	7

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

In the pattern where the mean of each treatment group is equal to the other at each center (Pattern 11), the error rate of the test is near the appropriate error level of 2.5%. For Pattern 10, the power of the test to show that neither treatment is equal to or superior to the other at all three centers is 7% for Treatment 1 and 29% for Treatment 2. The difference in these two results is due to the unequal sample size at the two centers where the treatment means are equal in magnitude but differ in signs.

For all patterns other than Patterns 11 and 10, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 ranges from 49% to 95%. For Patterns 1, 2, 3, 4, and 5 (patterns with no interaction or quantitative interaction), the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at all three centers is less than or equal to 1%. However, for the patterns with qualitative interaction (Patterns 6, 7, 8, 9 and 10), the power to detect that Treatment 2 is not \geq Treatment 1 at all three centers never exceeds 29%.

7.6.4 Two-stage Test Results

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 71.

There are notable differences in the means from the two models for Treatment 2, only the means for Patterns 1 and 11 do not differ substantially. The smaller than expected Type II means for these patterns reflect the relatively smaller sample size at Center 3, where Treatment 2 means are greater than the respective Treatment 2 means at one or both of the other centers.

The tests of significance of the difference between the two treatment means are consistent with the results of the calculation of the treatment means for the two models. The test results for the two models are only consistent for Pattern 11.

Table 71

Treatment Means for Each Treatment Calculated From Type II and Type III
Analyses and the Percentages of Simulations With Significant Tests
That Treatment 2 is Not Equal to Treatment 1 for Case 3.5

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III	Type II	Type III	Type II	Type III	Type II
	Model	Model				
1 NI	60	79	-0.009	-0.002	0.499	0.499
2 QN	60	36	-0.009	-0.002	0.499	0.284
3 QN	60	62	-0.009	-0.002	0.499	0.409
4 QN	60	43	-0.009	-0.002	0.499	0.320
5 QN	60	11	-0.009	-0.002	0.499	0.140
6 QL	60	39	-0.009	-0.002	0.499	0.320
7 QL	60	10	-0.009	-0.002	0.499	0.140
8 QL	60	3	-0.009	-0.002	0.499	-0.040
9 QL	60	3	-0.009	-0.002	0.499	-0.040
11 NI	5	4	-0.009	-0.002	-0.001	-0.001
10 QL	5	17	-0.009	-0.002	-0.001	-0.180

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For most patterns with an overall treatment effect and some form of interaction (all except Patterns 1, 10, 11), the number of significant test results is higher for the Type III model than for the Type II model. For all of these patterns, the Type II means for Treatment 2 are less than the respective Type III means. Pattern 3 is an exception; for

this pattern the power for the Type II and Type III models are nearly equal. The Type II mean is 0.409 versus the Type III mean of 0.499.

For Pattern 10, with interaction but no treatment difference (in the Type III model), the number of significant tests is much higher for the Type II model than for the Type III model. The results of the test reflect the smaller sample size at Center 3, where the mean of Treatment 2 is greater than the mean of Treatment 1.

Table 72 summarizes test results when a two-stage testing procedure is used to select the final analysis model.

For patterns with no interaction (Patterns 1 and 11), the percentages of differences found to be significant did not differ substantially depending on the method used as a preliminary test.

For patterns with a significant treatment effect and interaction (all patterns except Patterns 1, 11, 10), the power to detect the difference was generally highest if one of the tests for overall interaction was used as a preliminary test. (Patterns 3 and 6 were exceptions.) Using these preliminary tests, the power to detect treatment differences ranged from 54% to 60%. For Patterns 3 and 6, the tests for overall interaction and the methods for detecting qualitative interaction gave similar results. Of the tests for qualitative interaction, the exact method of Azzalini and Cox generally provided the best power. For patterns with quantitative interaction (Patterns 2, 3, 4 and 5), the range of the differences among the methods of detecting quantitative interaction for a specific pattern was less than 8%. For these patterns, power ranged from 13% to 65%. For the patterns with qualitative interaction (Patterns 6, 7, 8 and 9), the range was 8% to 21% within a pattern and the method providing the best power varied among the patterns. The power across all of these patterns ranged from 19% to 61%.

The power for Pattern 1 is at or near the expectation of 80% for all patterns. The results for Pattern 11 reflect the expected error rate of 5%. The error rate for Pattern 10 is higher (11% to 16%) due to the relatively smaller sample size at Center 3.

Table 72

Percentage of Simulations With Significant Test That Treatment 2 is Not Equal to Treatment 1 After Model Selection With a Preliminary Test of Interaction for Case 3.5

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	77	77	78	77	79	78	69
2 QN	54	54	38	36	36	37	49
3 QN	64	64	65	62	63	62	64
4 QN	58	59	47	44	44	43	58
5 QN	56	56	20	13	13	21	49
6 QL	60	60	61	58	58	53	61
7 QL	60	60	52	40	40	31	60
8 QL	60	60	47	33	34	37	60
9 QL	60	60	36	19	22	40	58
11 NI	6	6	4	4	4	4	5
10 QL	12	12	14	11	16	13	8

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

7.7 Case 3.6: 29 Patients at Center 1, 6 Patients at Center 2 and 29 Patients at Center 3

The simulations for this unequal sample size configuration generated 29 observations for each treatment group at Center 1, 6 observations for each treatment group at Center 2 and 29 observations for each treatment group at Center 3. For this case, the evidence of qualitative interaction should be strong for Patterns 6, 7 and 8, for which Center 1 has a negative treatment effect and Center 3 has a positive treatment effect. For Pattern 9, Centers 1 and 2 both have weak negative effects; hence the evidence is weaker.

7.7.1 Tests of Overall Interaction

The percentage of significant results for each test from the 2500 simulated datasets of each interaction pattern is presented in Table 73. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

The two tests provide results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in 99% or more of the individual simulations for each pattern.

The power of the tests to detect interaction in the patterns with qualitative interaction and strong quantitative interaction when there is a significant overall treatment difference (Patterns 5, 6, 7, 8 and 9) is near 100%. The power to detect patterns of quantitative interaction (Pattern 2, 3, 4 and 5) ranges from 57% to 99%. In the patterns with no interaction (Patterns 1 and 11), the error rate for both methods is near the expected error rate of 10%. The power to detect the interaction when there is no overall significant treatment difference (Pattern 10) is near 80%.

Table 73

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.6

Pattern and Int. Type*	Characteristics of Simulation Patterns					Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Center 3 Effect	Analysis of Variance	G. and S. H Statistic
1 NI	Yes	None	0.5	0.5	0.5	9	9
2 QN	Yes	Quant.	0.2	0.2	1.1	71	71
3 QN	Yes	Quant.	0.0	0.75	0.75	57	58
4 QN	Yes	Quant.	0.0	0.5	1.0	76	77
5 QN	Yes	Quant.	0.0	0.0	1.5	99	99
6 QL	Yes	Qual.	-0.5	1.0	1.0	99	99
7 QL	Yes	Qual.	-0.5	0.5	1.5	100	100
8 QL	Yes	Qual.	-0.5	0.0	2.0	100	100
9 QL	Yes	Qual.	-0.25	-0.25	2.0	100	100
11 NI	No	None	0.0	0.0	0.0	9	9
10 QL	No	Qual.	-0.5	0.0	0.5	76	77

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

7.7.2 Methods for the Detection of Qualitative Interaction

The percentages of simulated datasets with qualitative interaction, as evaluated using the tests of Azzalini and Cox (exact and approximate methods) and Gail and Simon, and the method of Ciminera et al., are presented in Table 74, along with the percentages of datasets with differing signs in the raw data.

Table 74

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.6

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	2	4	1	3	23
2 QN	9	14	3	13	50
3 QN	13	5	4	5	56
4 QN	15	9	4	12	60
5 QN	25	25	9	33	75
6 QL	80	61	59	44	98
7 QL	80	63	60	52	98
8 QL	83	69	64	70	99
9 QL	57	51	36	64	95
11 NI	9	4	1	8	76
10 QL	66	45	40	41	97

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The results from the exact and approximate approaches of Azzalini and Cox are different in this case. The exact approach generally has a higher rejection rate than the approximate approach.

The three methods provide similar results in Pattern 1. However, test results differ markedly among the three for the other patterns. The presence of differing signs

in the raw data consistently has a higher rejection rate than any of the three proposed methods.

For the patterns with quantitative interaction (Patterns 2, 3, 4 and 5), error rates were below 10% for the test of Gail and Simon. For Patterns 6, 7 and 8, the number of detected qualitative interactions is generally lowest for the method of Ciminera et al. (44% to 70%), consistently highest for the exact method of Azzalini and Cox (80% to 83%) and intermediate for the test of Gail and Simon ((59% to 64%) and the approximate method of Azzalini and Cox (61% to 69%). For Pattern 9, the number of detected qualitative interactions is highest for the method of Ciminera et al. and the exact method of Azzalini and Cox (64%, 57%). The test of Gail and Simon and the approximate method of Azzalini and Cox are lower (36%, 51%).

In Pattern 11, the error rates of the exact method of Azzalini and Cox and the method of Ciminera et al. are highest (9%, 8%), but seem appropriate for a test with a level of 10%. The error rates of the approximate method of Azzalini and Cox and the Gail and Simon test are much less (4%, 1%). For Pattern 10, the detection rate for the exact method of Azzalini and Cox is 66%, and that for the other methods is near 40%.

7.7.3 Test of Non-inferiority

The percentages of significant tests of the null hypotheses that each respective treatment group is at least as good as the other treatment group at every center, tested using the test of non-inferiority proposed by Gail and Simon, are presented in Table 75.

Table 75

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.6

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at All Centers	Treatment 1 is not \geq Treatment 2 at All Centers
1 NI	0	71
2 QN	0	96
3 QN	0	78
4 QN	0	93
5 QN	1	100
6 QL	24	95
7 QL	25	100
8 QL	28	100
9 QL	9	100
11 NI	3	3
10 QL	28	27

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

In Pattern 11, the error rate of the test is near the appropriate error level of 2.5%. For Pattern 10, the power of the test to show that neither treatment is equal to or superior to the other at all three centers is 28% for Treatment 1 or 27% for Treatment 2, respectively.

For all patterns other than Patterns 11 and 10, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 is at least 71%. For Patterns

1, 2, 3, 4, and 5 (patterns with no interaction or quantitative interaction), the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at all three centers is less than or equal to 1%. However, for the patterns with qualitative interaction (Patterns 6, 7, 8, 9 and 10), the power to detect that Treatment 2 is not \geq Treatment 1 at all three centers never exceeds 28%.

7.7.4 Two-stage Test Results

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 76.

There are notable differences in the means from the two models for Treatment 2, although the means for several patterns (Patterns 1, 4, 7, 10 and 11) do not differ substantially. The differences between the Type II means and simulation target means for the other patterns reflect the relatively smaller sample size at Center 2. For Patterns 3 and 6, the Treatment 2 means are smaller than expected and for Patterns 2, 5, 8 and 9, Treatment 2 means are greater than the expected.

The tests of significance of the difference between the two treatment means are generally consistent with the results of the calculation of the treatment means for the two models. For patterns with smaller than expected Type II means for Treatment 2 (Patterns 3 and 6), the Type III test has power equal to or greater than the Type II test. For patterns with larger than expected Type II means (Patterns 2, 5, 8 and 9), the Type II test is more powerful. The Type II model is also more powerful for patterns with an overall treatment effect and similar Type III and Type II means (Patterns 1, 4, 7), For Patterns 11 and 10, the test results for the two models are consistent.

Table 76

Treatment Means for Each Treatment Calculated From Type II and Type III
Analyses and the Percentages of Simulations With Significant Tests
That Treatment 2 is Not Equal to Treatment 1 for Case 3.6

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III	Type II	Type III	Type II	Type III	Type II
	Model	Model				
1 NI	57	80	0.001	0.001	0.498	0.496
2 QN	57	92	0.001	0.001	0.498	0.604
3 QN	57	61	0.001	0.001	0.498	0.407
4 QN	57	79	0.001	0.001	0.498	0.496
5 QN	57	96	0.001	0.001	0.498	0.676
6 QL	57	36	0.001	0.001	0.498	0.317
7 QL	57	74	0.001	0.001	0.498	0.496
8 QL	57	94	0.001	0.001	0.498	0.676
9 QL	57	98	0.001	0.001	0.498	0.766
11 NI	5	5	0.001	0.001	-0.002	-0.004
10 QL	5	5	0.001	0.001	-0.002	-0.004

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

Table 77 summarizes test results when a two-stage testing procedure is used to select the final analysis model.

Table 77

Percentage of Simulations With Significant Test That Treatment 2
is Not Equal to Treatment 1 After Model Selection With
a Preliminary Test of Interaction for Case 3.6

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	78	78	79	77	80	79	69
2 QN	65	65	86	82	91	83	63
3 QN	62	62	63	61	61	60	61
4 QN	63	62	78	76	79	74	63
5 QN	58	58	79	77	90	70	58
6 QL	58	58	58	55	54	49	58
7 QL	57	57	62	65	66	64	58
8 QL	57	57	60	61	65	59	57
9 QL	57	57	64	64	73	59	57
11 NI	7	7	5	5	5	5	5
10 QL	6	6	7	6	6	5	6

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For patterns with no interaction and/or no treatment effect (Patterns 1, 11 and 10), the percentages of differences found to be significant did not differ substantially depending on the method used as a preliminary test. For Patterns 3, 6, 7 and 8 the difference in the power of the methods varied less than 10% for a given pattern.

Generally, the overall tests of interaction provided less power than the methods of detecting quantitative interaction.

For most of the patterns, the differences among the methods of detecting quantitative interaction were small. The differences among the methods within a given pattern only exceeded 10% for Patterns 5 and 9. For six of the eight patterns with a significant treatment effect and interaction (all patterns except Patterns 1, 11, 10), the test of Gail and Simon had the most power to detect the difference (54% to 91%).

The power for Pattern 1 is at or near the expectation of 80% for all patterns. The results for Patterns 11 and 10 reflect the expected error rate of 5%.

7.8 Case 3.7: 6 Patients at Center 1, 29 Patients at Center 2 and 29 Patients at Center 3

The simulations for this unequal sample size configuration generated 6 observations for each treatment group at Center 1, 29 observations for each treatment group at Center 2 and 29 observations for each treatment group at Center 3. For this case, the evidence of qualitative interaction should not be strong for Patterns 6, 7 and 8, for which Center 2 has a positive or null treatment effect and Center 3 has a positive treatment effect. For Pattern 9, Centers 1 and 2 both have weak negative effects; hence the evidence is stronger.

7.8.1 Tests of Overall Interaction

The percentage of significant results for each test from the 2500 simulated datasets of each interaction pattern is presented in Table 78. Also presented in the table are several distinguishing characteristics of the simulated interaction patterns.

Table 78

Characteristics of Simulation Patterns and Percentage of Simulations With Significant Tests of Overall Interaction for Case 3.7

Pattern and Int. Type*	Characteristics of Simulation Patterns					Interaction Tests	
	Overall Treatment Effect	Interaction Type	Center 1 Effect	Center 2 Effect	Center 3 Effect	Analysis of Variance	G. and S. H Statistic
1 NI	Yes	None	0.5	0.5	0.5	10	10
2 QN	Yes	Quant.	0.2	0.2	1.1	72	72
3 QN	Yes	Quant.	0.0	0.75	0.75	27	28
4 QN	Yes	Quant.	0.0	0.5	1.0	46	47
5 QN	Yes	Quant.	0.0	0.0	1.5	99	99
6 QL	Yes	Qual.	-0.5	1.0	1.0	71	72
7 QL	Yes	Qual.	-0.5	0.5	1.5	95	95
8 QL	Yes	Qual.	-0.5	0.0	2.0	100	100
9 QL	Yes	Qual.	-0.25	-0.25	2.0	100	100
11 NI	No	None	0.0	0.0	0.0	10	10
10 QL	No	Qual.	-0.5	0.0	0.5	46	47

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The two tests provide results that are very similar. An examination of the individual test results of the 2500 simulations for each pattern indicates that the two tests are in agreement in 99% or more of the individual simulations for each pattern.

The power of the tests to detect interaction in the patterns with qualitative interaction (Patterns 6, 7, 8 and 9) is equal to or greater than 72%. The power to

detect patterns of quantitative interaction (Pattern 2, 3, 4 and 5) ranges from 27% to 99%. In the patterns with no interaction (Patterns 1 and 11), the error rate for both methods is near the expected error rate of 10%.

7.8.2 Methods for the Detection of Qualitative Interaction

The percentages of simulated datasets with qualitative interaction, as evaluated using the tests of Azzalini and Cox (exact and approximate methods) and Gail and Simon, and the method of Ciminera et al., are presented in Table 79, along with the percentages of datasets with differing signs in the raw data.

The results from the exact and approximate approaches of Azzalini and Cox are different in this case. The exact approach generally has a lower rejection rate than the approximate approach.

The three methods provide similar results in the pattern with no interaction and an overall treatment difference (Pattern 1). For the other patterns with overall treatment differences (all other patterns except Patterns 10 and 11), the number of detected qualitative interactions is consistently lowest for the test of Gail and Simon. The presence of differing signs in the raw data always has a higher rejection rate than any of the three proposed methods.

For those patterns with an overall treatment effect and qualitative interaction (Patterns 6, 7, 8, and 9), the rates of detection of qualitative interaction ranged from 20% to 36% for the test of Gail and Simon, from 43% to 64% for the method of Ciminera et al. and 41% to 57% for the methods of Azzalini and Cox.

The error rates for the patterns with quantitative interaction (Patterns 2, 3, 4 and 5) are all 10% or lower for the test of Gail and Simon.

Table 79

Percentage of Simulations With Significant Tests or Substantial Evidence of Qualitative Interaction for Case 3.7

Pattern and Int. Type*	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
1 NI	2	5	0	3	23
2 QN	10	14	3	14	51
3 QN	13	22	4	15	50
4 QN	13	22	5	15	51
5 QN	24	25	10	33	75
6 QL	41	55	20	43	80
7 QL	41	55	20	43	80
8 QL	48	56	26	61	90
9 QL	57	52	36	64	94
11 NI	10	4	1	9	74
10 QL	38	32	16	37	89

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

In the pattern with no interaction, and no treatment effect (Pattern 11), the error rates for all of the methods are less than or equal to the expected level of 10%. For Pattern 10, the detection rate for Gail and Simon is 16%, and that for the other methods is greater than 32%.

7.8.3 Test of Non-inferiority

The percentages of significant tests of the null hypotheses that each respective treatment group is at least as good as the other treatment group at every center, tested using the test of non-inferiority proposed by Gail and Simon are presented in Table 80.

Table 80

Percentage of Simulations With a Significant Test That One Treatment Group Mean is Not \geq the Other Treatment Group Mean at All Centers for Case 3.7

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at All Centers	Treatment 1 is not \geq Treatment 2 at All Centers
1 NI	0	71
2 QN	0	96
3 QN	0	95
4 QN	0	96
5 QN	1	100
6 QL	5	100
7 QL	5	100
8 QL	7	100
9 QL	10	100
11 NI	3	2
10 QL	7	28

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

In the pattern where the mean of each treatment group is equal to the other at each center (Pattern 11), the error rate of the test is near the appropriate error level of 2.5%. For Pattern 10, the power of the test to show that neither treatment is equal to or superior to the other at all three centers is 28% for Treatment 1 and 7% for Treatment 2. The difference in these two results is due to the unequal sample size at the two centers where the treatment means are equal in magnitude but differ in signs.

For all patterns other than Patterns 11 and 10, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 is at least 71%. For Patterns 1, 2, 3, 4, and 5 (patterns with no interaction or quantitative interaction), the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at all three centers is less than or equal to 1%. However, for the patterns with qualitative interaction (Patterns 6, 7, 8, 9 and 10), the power to detect that Treatment 2 is not \geq Treatment 1 at all three centers never exceeds 10%.

7.8.4 Two-stage Test Results

The Type III and Type II treatment means and the results of analyses using Type III and Type II models are presented in Table 81.

There are notable differences in the means from the two models for Treatment 2, only the means for Patterns 1 and 11 do not differ substantially. The larger than expected Type II means for these patterns reflect the relatively smaller sample size at Center 1, where Treatment 2 means are less than the respective Treatment 2 means at one or both of the other centers.

Table 81

Treatment Means for Each Treatment Calculated From Type II and Type III
Analyses and the Percentages of Simulations With Significant Tests
That Treatment 2 is Not Equal to Treatment 1 for Case 3.7

Pattern and Int. Type*	Percentage of Simulations With Significant Tests		Treatment 1 Means		Treatment 2 Means	
	Type III	Type II	Type III	Type II	Type III	Type II
	Model	Model				
1 NI	57	80	0.001	0.001	0.498	0.497
2 QN	57	92	0.001	0.001	0.498	0.605
3 QN	57	97	0.001	0.001	0.498	0.677
4 QN	57	97	0.001	0.001	0.498	0.677
5 QN	57	96	0.001	0.001	0.498	0.677
6 QL	57	100	0.001	0.001	0.498	0.856
7 QL	57	100	0.001	0.001	0.498	0.856
8 QL	57	99	0.001	0.001	0.498	0.856
9 QL	57	98	0.001	0.001	0.498	0.766
11 NI	5	5	0.001	0.001	-0.002	-0.003
10 QL	5	16	0.001	0.001	-0.002	0.177

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

The tests of significance of the difference between the two treatment means are consistent with the results of the calculation of the treatment means for the two models. The test results for the two models are only consistent for Pattern 11. For all patterns except Pattern 11, the number of significant test results is higher for the Type

II model than for the Type III model. For all of these patterns except Pattern 1, the Type II means for Treatment 2 are greater than the respective Type III means.

Table 82 summarizes test results when a two-stage testing procedure is used to select the final analysis model.

For patterns with no interaction (Patterns 1 and 11), the percentages of differences found to be significant did not differ substantially depending on the method used as a preliminary test.

For those patterns with a significant treatment effect and interaction (all patterns except Patterns 1, 11, 10), the power to detect the difference was highest if the test of Gail and Simon was used as a preliminary test. Generally, the power of the methods for the detection of qualitative interaction exceeded that of the tests of overall interaction. Using the tests of overall interaction as preliminary tests, the power to detect treatment differences ranged from 57% to 77%. Of the tests for qualitative interaction, the test of Gail and Simon consistently provided the best power, from 74% to 93%. The power for the methods of Azzalini and Cox and Ciminera et al. ranged from 58% to 85% for these patterns. The differences among methods ranged from 9% to 20%.

The power for Pattern 1 is at or near the expectation of 80% for all patterns. The results for Pattern 11 reflect the expected error rate of 5%. The error rate for Pattern 10 is higher (12% to 16%) due to the relatively smaller sample size at Center 1.

Table 82

Percentage of Simulations With Significant Test That Treatment 2
is Not Equal to Treatment 1 After Model Selection With
a Preliminary Test of Interaction for Case 3.7

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	78	77	79	77	80	78	69
2 QN	64	64	85	81	90	82	62
3 QN	78	77	85	78	93	84	62
4 QN	72	71	85	78	93	83	62
5 QN	57	57	79	76	89	70	58
6 QL	61	62	67	61	81	66	57
7 QL	58	58	67	61	81	66	57
8 QL	57	57	64	61	76	59	57
9 QL	57	57	64	63	74	58	57
11 NI	6	6	5	5	5	5	5
10 QL	12	12	15	12	16	13	8

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

7.9 Discussion

7.9.1 Tests of Overall Interaction and Methods for the Detection of Qualitative Interaction

It is clear from the results presented in this chapter that the reliability of the tests of qualitative interaction to accurately detect qualitative interaction depends both on the degree of interaction present in the data and the relative sample sizes at the three centers. For the cases presented in this study, the minimum and maximum percentages of simulations with significant tests of overall interaction and significant tests or substantial evidence of qualitative interaction are presented in Table 83.

A comparison of the results of Pattern 1 across the seven sample size configurations provides an assessment of the methods when there is an overall treatment difference and no interaction. The error rate for the two tests of overall interaction is near the expected rate of 10% for all sample size configurations. The error rates for the tests of qualitative interaction range from 0% to 5%. The percentage of datasets with treatment effect reversals in the raw data ranges from 14% to 24% for this pattern.

The results for Pattern 11 are generally very consistent across the different sample size cases. The results for the test of overall interaction and the exact method of Azzalini and Cox are all from 8% to 11%. The approximate method of Azzalini and Cox ranges from 4% to 8% and the test of Gail and Simon from 1% to 2%. The method of Ciminera et al. has the largest range for this pattern, from 4% to 18%.

For Pattern 10, the method of Ciminera et al. has the narrowest range of the methods, from 37% to 51%. For the other methods, the widths of the ranges are similar from 23 to 32 percentage points, but the lower and upper limits vary. The

tests of overall interaction range from 45% to 77%, Azzalini and Cox from 37% to 66% and 32% to 55%, and Gail and Simon from 15% to 40%.

Table 83

Minimum and Maximum Percentages of Simulations With Significant Tests or Substantial Evidence of Overall Interaction or Qualitative Interaction for Three-Center Simulations

Pattern and Int. Type*	Overall Interaction		Qualitative Interaction				
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	8, 11	9, 11	1, 2	1, 5	0, 1	0, 3	14, 24
2 QN	37, 72	38, 72	6, 10	1, 14	1, 3	5, 20	40, 53
3 QN	27, 58	28, 59	12, 13	4, 22	3, 5	3, 47	50, 56
4 QN	44, 76	45, 77	12, 15	4, 22	3, 5	4, 47	51, 60
5 QN	71, 99	72, 99	23, 25	8, 30	8, 10	17, 53	74, 76
6 QL	71, 99	72, 99	41, 86	55, 70	20, 68	40, 98	80, 99
7 QL	93, 100	94, 100	41, 86	55, 70	20, 69	42, 99	80, 99
8 QL	97, 100	97, 100	48, 88	56, 75	26, 74	52, 99	90, 99
9 QL	96, 100	96, 100	54, 67	40, 61	30, 45	45, 90	93, 97
11 NI	8, 11	9, 11	8, 10	4, 8	1, 2	4, 18	74, 76
10 QL	44, 76	45, 77	37, 66	32, 55	15, 40	37, 51	89, 97

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

In the presence of an overall treatment difference (all patterns except Patterns 11 and 10), the use of the differing signs among the raw means as an indication of the

presence of qualitative interaction generally gives a result that is intermediate between the higher tests of overall interaction and the lower methods for the detection of qualitative interaction. For Patterns 11 and 10, the percentages are much higher when the raw means are used to detect qualitative interaction.

Patterns 2, 3 and 4 have an overall treatment difference and moderate quantitative interaction. There are positive treatment differences at two of the three centers and either a positive or null difference at the third. When sample sizes are equal, the tests of overall interaction are significant in 53% to 66% of the simulations (Table 48). In the cases of unequal sample sizes, the percentages are generally lower but vary from 27% to 77%. The error rates of the methods for detecting quantitative interaction are generally very low for these patterns, even in the cases with unequal sample sizes. The test of Gail and Simon is significant in less than 5% of the simulations for these patterns. The percentage of the significant results for the exact method of Azzalini and Cox ranges from 6% to 15%, while the results for the approximate method range from 1% to 22%. However, the method of Ciminera et al. is more variable: the range is from 5% to 47%. The highest error rates occur in Case 3.2, which has a sample size ratio of 39:13:12.

Pattern 5 has one center with a positive difference and two centers with no between treatment difference. In the case of equal sample sizes, the tests of overall interaction are significant for 98% of the simulations (Table 48), Gail and Simon has an error rate of 10% and the other quantitative interaction methods vary from 24% to 27% (Table 49). The percentages of significant results for the tests of overall interaction vary from 71% to 99% for the cases with unequal sample sizes. The results for the test of Gail and Simon are very consistent (8% to 10%), as are the exact Azzalini and Cox results (23% to 25%). The percentage of detected qualitative

interactions for the approximate method of Azzalini and Cox ranges from 8% to 30% and the range of the Ciminera et al. method is 17% to 53%.

Patterns 6, 7, 8 and 9 have overall treatment differences and qualitative interactions. The tests of overall interaction have high percentages of significant results for all of these patterns for all sample size patterns. The percentages of significant overall tests are 98% or greater for all patterns when sample sizes are equal (Table 48). In the cases of unequal sample sizes, the percentages vary from 93% to 100%, except for Pattern 6. The percentages for this pattern, which has relatively large positive differences at Centers 2 and 3 and a negative, clinically significant difference at Center 1, vary from 71% to 99%.

For these four patterns displaying qualitative interaction, the results for the methods of detecting qualitative interaction can be grouped by pattern. Patterns 6, 7 and 8 all have large positive differences at Center 3 and negative differences equal to the clinically significant difference at Center 1. Center 2 is either positive or neutral for these patterns. The relationship between sample size configuration and the importance of the respective qualitative interaction is generally similar for these three patterns. Pattern 9, with a large positive difference at Center 3 and two relatively small negative differences at Centers 1 and 2 has a different behavior.

For the methods of detecting qualitative interaction, the results for Patterns 6, 7 and 8 are generally very similar when sample sizes are equal (Table 49). The percentages for Azzalini and Cox vary from 70% to 75% and from 47% to 54% for Gail and Simon. The results for Ciminera et al. are more variable, from 40% to 69%. For Pattern 9, the results for Azzalini and Cox (62%, 61%) and Gail and Simon (39%) are slightly lower than those for the other patterns. However, the method of Ciminera et al. finds “substantial evidence” in 70% of the simulations.

Sample size has an important effect on the percentages of qualitative interactions detected for these patterns. The combination of treatment differences and samples sizes create scenarios with stronger qualitative interaction and others with weaker interaction. For Patterns 6, 7 and 8, Cases 3.2, 3.5 and 3.6 have the strongest qualitative interaction, while the qualitative interaction in Cases 3.3, 3.4 and 3.7 is relatively weaker. In the first group, sample sizes are large at the negative Center 1; while in the second group, sample sizes at this center are smaller. The characteristics of these “strong” and “weak” cases of qualitative interaction are summarized in Table 84.

Table 85 shows the differences between the percentages of significant simulations detected for the “strong” cases and those for the “weak” cases.

The approximate method of Azzalini and Cox shows the least degree of discrimination, with a difference between the “strong” and “weak” patterns of only 5% to 7%. The method of Ciminera et al. is better, with a range of differences from 18% to 21%. The exact method of Azzalini and Cox (27% to 31%) and the test of Gail and Simon (31% to 33%) have the most power to differentiate between the patterns. For the “strong” patterns, the exact method of Azzalini and Cox has the highest mean percentage of detected simulations (83%), while that of the other methods ranges from 64% to 68%. For the “weak” patterns, the test of Gail and Simon has an average percentage of 31% and that of the other methods ranges from 49% to 60%.

The use of differing signs in the raw data lacks the ability to discriminate between the strong and the weak. The average percentage of simulations with differing signs is 89% in the weak cases and 98% in the strong cases.

Table 84

Sample Sizes by Center for “Strong” and “Weak” Cases of Qualitative Interaction. Effect Sizes for Each Pattern at Each Center are Also Indicated.

Group	Characteristic	Center 1			Center 2			Center 3		
	Pattern	6	7	8	6	7	8	6	7	8
	Effect size	-0.5	-0.5	-0.5	1.0	0.5	0.0	1.0	1.5	2.0
Strong	Case 3.2	39			13			12		
Strong	Case 3.5	29			29			6		
Strong	Case 3.6	29			6			29		
Weak	Case 3.3	13			39			12		
Weak	Case 3.4	13			12			39		
Weak	Case 3.7	6			29			29		

For Pattern 9, the influence of sample size is generally less pronounced than in the other patterns with qualitative interaction (Table 83). The ranges for the patterns are generally narrower, except for Ciminera et al., which is from 45% to 90%. The ranges for the other methods are: exact Azzalini and Cox, 54% to 67%; approximate Azzalini and Cox, 40% to 61%; Gail and Simon, 30% to 45%.

7.9.2 Test of Non-inferiority

The minimum and maximum percentages of the significant tests of non-inferiority, tested using the test of non-inferiority proposed by Gail and Simon, are presented in Table 86.

Table 85

Differences in Qualitative Interactions Detected in Cases of “Strong” Qualitative Interaction and Cases of “Weak” Qualitative Interaction for Three-Center Simulated Data

Pattern	Percentage With Significant Tests			Percentage With Substantial Evidence	
	Azzalini and Cox Exact	Azzalini and Cox Approx.	Gail and Simon	Ciminera et al.	Raw Data
6	31	5	33	20	12
7	30	5	33	21	11
8	27	7	31	18	6

The error rate of the test is near the appropriate error level of 2.5% for Pattern 11, where the mean of each treatment group is equal to the other at each center. This result is consistent across all sample size cases. Pattern 11 is the only pattern where, for at least one of the centers, Treatment 2 is not numerically superior to Treatment 1 by at least 0.5, which represents the clinically relevant difference the study is powered to detect. For Pattern 10, the power of the test to show that neither treatment is equal to or superior to the other at all three centers is very dependent on the sample size configuration. The power ranges from 7% to 38% for both Treatment 1 and Treatment 2. These results correspond with the unequal sample sizes at the two centers where the treatment means are equal in magnitude but differ in signs.

For all patterns other than Patterns 11 and 10, the power to detect that Treatment 1 is not equal to or greater than Treatment 2 is generally at least 71%.

Table 86

**Minimum and Maximum Percentages of Simulations With a Significant Test That
One Treatment Group Mean is Not \geq the Other Treatment Group Mean
at Both Centers**

Pattern and Int. Type*	Treatment 2 is not \geq Treatment 1 at All Centers	Treatment 1 is not \geq Treatment 2 at All Centers
1 NI	0, 0	71, 71
2 QN	0, 0	49, 99
3 QN	0, 1	61, 95
4 QN	0, 1	62, 98
5 QN	1, 2	59, 100
6 QL	5, 33	85, 100
7 QL	5, 33	79, 100
8 QL	7, 37	82, 100
9 QL	8, 16	81, 100
11 NI	2, 3	2, 3
10 QL	7, 38	7, 38

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

For Patterns 1, 2, 3, 4, and 5 (patterns with no interaction or quantitative interaction), the error rate of not finding Treatment 2 to be equal to or greater than Treatment 1 at all three centers is always less than or equal to 2%. However, for the patterns with qualitative interaction (Patterns 6, 7, 8, 9 and 10), the power to detect that Treatment 2 is not \geq Treatment 1 at all three centers never exceeds 37%. This percentage is highly related to the sample size configuration. For the cases of “weak” qualitative

interaction for Patterns 6, 7 and 8 as described in the previous section, the power ranges from 5% to 12%. For the cases of “strong” interaction, the range is from 24% to 37%.

7.9.3 Two-stage Test Results

The ranges of significant test results for the test of difference between the two treatment groups, when a two-stage testing procedure is used to select the final analysis model, are presented in Table 87.

A comparison of the results of Pattern 1 across the seven sample size configurations provides an assessment of the methods when there is a significant treatment difference and no significant interaction. For Pattern 1, the power to detect the treatment difference through a two-stage testing procedure, using any of the tests of overall interaction or qualitative interaction as a pretest, is always as powerful as using Type III model. As the imbalance increases, the power of the Type III model decreases, whereas that of a model using any of the pretest methods remains near 80%. In the cases with greatest degree of imbalance, the power with a Type III model may be as low as 57% for the cases simulated in this study, whereas using a two-stage procedure provides power greater than 77% (Table 81 and Table 82).

In the discussion of the case of equal sample sizes (Table 52), we noted that for patterns with interaction the power was near 80% using any of the methods as a pretest. For the patterns with no interaction, the error rate was at the expected rate of 5% for all methods.

Table 87

**Minimum and Maximum Percentages of Simulations With Significant Tests
That Treatment 2 is Not Equal to Treatment 1 After Model Selection
With a Preliminary Test of Interaction**

Pattern and Int. Type*	Method Used for Preliminary Test						
	Analysis of Variance	Gail and Simon H	Azz. and Cox Exact	Azz. and Cox Approx.	Gail and Simon	Cim. et al.	Raw Data
1 NI	77, 81	77, 81	78, 80	77, 80	79, 80	78, 81	69, 81
2 QN	54, 80	54, 80	38, 91	36, 89	36, 96	37, 94	49, 80
3 QN	60, 80	61, 80	41, 85	37, 83	37, 93	59, 87	59, 80
4 QN	58, 80	59, 80	40, 88	36, 86	36, 94	43, 91	58, 80
5 QN	56, 80	56, 80	20, 82	13, 79	13, 91	21, 86	49, 80
6 QL	58, 80	58, 80	58, 80	47, 80	47, 81	49, 79	57, 80
7 QL	57, 80	57, 80	52, 80	40, 80	40, 81	31, 78	57, 80
8 QL	57, 80	57, 80	47, 80	33, 80	34, 79	37, 80	57, 80
9 QL	57, 80	57, 80	36, 79	19, 79	22, 78	40, 80	57, 80
11 NI	5, 7	5, 7	4, 6	4, 6	4, 6	4, 7	4, 6
10 QL	5, 12	5, 12	5, 16	5, 14	5, 19	5, 19	5, 8

*Interaction Type: NI = No Interaction, QN = Quantitative Int., QL = Qualitative Int.

A more general evaluation of the two-stage testing procedure incorporating the methods of detecting qualitative interaction is not appropriate for the other cases of three center simulations. The simulations presented in this study do not provide an accurate assessment of the method. The Type II treatment means are highly influenced by the respective treatment pattern and sample size. The power of the

Type II analysis to detect between-treatment differences is a function of these treatment differences. And this power contributes to the power of the two-stage process through the selection of the Type II model by the first-stage test.

For example, in Case 3.3 Pattern 6, the Type II means are 0.000 and 0.693 for Treatments 1 and 2, respectively and the Type III means are 0.000 and 0.496 for Treatments 1 and 2, respectively (Table 61). The power of the test of differences between the two means is 97% for Type II and 66% for Type III. A two-stage procedure using the Gail and Simon procedure, which identifies significant qualitative interaction in 34% of the simulations (Table 59) has a power to detect significant treatment differences of 78% (Table 62). And in Case 3.6 Pattern 6, the Type II means are 0.001 and 0.317 for Treatments 1 and 2, respectively and the Type III means are 0.001 and 0.498 for Treatments 1 and 2, respectively (Table 76). The power of the test of differences between the two means is 36% for Type II and 57% for Type III. A two-stage procedure using the Gail and Simon procedure, which identifies significant qualitative interaction in 59% of the simulations (Table 74) has a power to detect significant treatment differences of 54% (Table 77).

CHAPTER VIII

CONCLUSIONS AND FUTURE WORK

8.1 Conclusions

The interpretation of positive results in a multicenter trial may be complicated by the presence of a significant treatment by center interaction. This confusion may be unnecessary when the interaction is quantitative and does not contradict an overall positive treatment difference. This study evaluates and compares three methods proposed for the detection of qualitative interaction.

The objective of this research is to evaluate and compare three methods for the detection of qualitative interaction proposed by Azzalini and Cox (1984), Gail and Simon (1985) and Ciminera et al. (1993). The methodology employed is the analysis of simulated data for multicenter studies of two and three centers with two treatments. The effect of unequal sample size and the presence of an overall treatment effect on the characteristics of the methods were examined. These methods are compared to a common, ad-hoc method of identifying qualitative interaction, assessing the signs of the treatment effects by center. Two tests of overall interaction: the ANOVA test of interaction and the H statistic proposed by Gail and Simon were also examined.

This study also evaluates the use of the methods of detecting qualitative interaction and the tests of overall interaction in a two-stage testing system. A test of non-inferiority related to one of the tests of qualitative interaction is also assessed.

The results of the present study show that the use of an overall test of interaction or the use of the presence of differing signs among the raw means to

determine the presence of contradictory interaction produces an overabundance of significant results. Pattern 4 in Case 3.7 provides an appropriate example (Table 78). Center 3 has 29 patients per treatment group with an average treatment effect of 1.0, Center 2 has 29 patients per group with an average treatment effect of 0.5 (the clinically significant difference) and Center 1 has 6 patients with no average treatment effect. The test for an overall treatment by center interaction is significant in 46% of the simulations (Table 78) and the raw means show an effect reversal in 51% of the simulations (Table 79). The methods proposed for the detection of qualitative interaction all present a more reasonable picture. The percentage of significant interactions detected by these methods range from 5% for the test of Gail and Simon to 22% for the exact method of Azzalini and Cox (Table 79).

The comparisons of the analysis of variance test of overall interaction with the Gail and Simon H test show the two tests produce equivalent results in 99% or more of the simulations for most patterns and sample size cases examined in this study.

The simulations of the test of non-inferiority proposed by Gail and Simon indicate that the test does not provide adequate power to reject the null hypothesis of non-inferiority in the presence of qualitative interaction. An example is Pattern 8 for Case 3.6 (Table 73 and Table 75). Center 1 has 29 patients per treatment group with a treatment difference of -0.5 (i.e. Treatment 2 < Treatment 1), Center 2 has 6 patients per group with no treatment effect and Center 3 has 29 patients per group with a treatment difference of 2.0 (i.e. Treatment 2 > Treatment 1). The test rejects the null hypothesis that Treatment 2 is not = Treatment 1, at a level of 0.25, in only 28% of the simulations.

The exact method of Azzalini and Cox outperforms the approximate method of the test in the simulations of this study. This is especially evident in the three-

center simulations. In the patterns with quantitative interaction, the approximate method is much more sensitive to sample size and shows a much higher error rate for some cases (see Table 79 [Case 3.7]), although the average error rate for all simulations is comparable for the two methods. In the patterns of qualitative interaction, the approximate method is less sensitive to differences between “strong” qualitative interaction and “weak” qualitative interaction (Table 85). This method has lower detection rates for “strong” patterns and higher detection rates for “weak” patterns than the exact method.

The test for qualitative interaction proposed by Gail and Simon is the recommended method for detecting qualitative interaction. The error rates for patterns with quantitative interaction are consistently lowest for this method among the compared methods for both two-center and three-center simulations. The error rate for patterns at the boundary of quantitative and qualitative interaction (e.g. two-center Pattern 2) is near the expected error rate of 10%. The sensitivity of the method to qualitative interaction is displayed in the comparison of “strong” and “weak” interaction cases for the three-center simulations.

The second choice would be the exact method of Azzalini and Cox, which has higher detection rates. The error rate for patterns with quantitative interaction is as high as 24% and 25% for two- and three-center simulations, respectively. However, this test is also more powerful than Gail and Simon in those patterns with qualitative interaction. The average power of 83% in the three-center patterns with “strong” interaction does not indicate that this test has an excessive detection rate. This test may be preferable to Gail and Simon in trials with small sample sizes.

This study does not provide a good evaluation of two-stage sampling, except for the cases with equal sample sizes. For the cases with equal sample sizes, two-

stage testing, with one of the methods recommended above, is preferable to using a Type II model. However, if the test for qualitative interaction is significant, the treatment-by-center cell means should be examined and the test for between – treatment differences may be un-informative.

A more general evaluation of the two-stage testing procedure incorporating the methods of detecting qualitative interaction is not appropriate for this study. The simulations presented do not provide an accurate assessment of the method. The Type II treatment means are highly influenced by the respective treatment pattern and sample size. The power of the Type II analysis to detect between-treatment differences is a function of these treatment differences. And this power contributes to the power of the two-stage process through the selection of the Type II model by the first-stage test.

This study was not designed to evaluate the powers of Type II and Type III models under conditions of unequal sample size. However, it is worth noting results that provide some comparison. The simulations of the study were designed to have 80% power in circumstances of equal sample size. For Pattern I (both two- and three-center), with treatment effects at all centers equal to the clinically significant difference, the power was as low as 52% in an extremely unbalanced design. This is only marginally better than the toss of an unbiased coin.

8.2 Future Work

The methods of Azzalini and Cox, Gail and Simon and Ciminera et al. are assessed under limited sets of circumstances. The recommendations made here are based on the assumptions of these simulations. Additional research under extended conditions would provide additional insight in to the value of these methods.

The method of analyzed and summarizing results of simulations used in this study is a valuable research tool and can be recommended as a method for use in future studies.

Some suggestions for future work are:

Analyze the error rates from this study with analysis of variance using, as factors in the model, the conditions of the simulation such as method, type and degree of interaction, and sample size.

Examine the methods for studies with additional centers. This study began with the most basic design. Generally, multicenter trials involve a larger number of centers.

Produce an alternative evaluation method for the two-stage testing procedure. The method used in this study did not provide an effective evaluation except for the condition of equal sample sizes.

Evaluate the methods of Gail and Simon under conditions of heterogeneous variances. These methods do not require the strict assumption of equal variance that we imposed upon them for this study.

Appendix
Simulation Programs

The production and analysis of the simulated datasets for this study was produced using SAS software. The master program for the two-center simulations was **m_master2.sas**. The master program for the three-center simulations was **m_master3.sas**. These programs called, directly or indirectly, the following SAS macros:

m_summary.sas

m_azzalini.sas

m_ciminera.sas

m_gailsimn.sas

m_recap.sas

m_saveperm.sas

m_listsave.sas

s2_sample_size.sas

s3_sample_size.sas

The results of the simulations and tests were summarized using the following programs, for two- and three-center simulations, respectively.

m_relate2.sas

m_relate3.sas

The two-stage testing procedure was examined using the following programs, for two- and three-center simulations, respectively.

m_ptest2.sas

m_ptest3.sas

All of these programs are presented in this appendix, in the same order that they are listed above.

```

*m_master2.sas      12may00;

*Simulation and Test Program for 2 centers;
*This program simulates data and analyzes it using the test macros;

*options ls = 78 ps = 52 pageno = 1 mprint mlogic symbolgen notes;
*options ls=78 ps=52 pageno=1 nomprint nomlogic nosymbolgen nosource;
options ls = 78 ps = 52 pageno = 1 nomprint nomlogic nosymbolgen nonotes;

%macro simulate;
data multi (keep = site trt npat x);
  seed2 = &seed + &nsim;
  do site = 1 to 2;
    if site = 1 then enrolled = &s1n;
    else if site = 2 then enrolled = &s2n;
    do trt = 1 to 2;
      do npat = 1 to enrolled;
        call rannor(seed2, x);
        if site = 1 and trt = 1 then x = x + &s1t1;
        if site = 1 and trt = 2 then x = x + &s1t2;
        if site = 2 and trt = 1 then x = x + &s2t1;
        if site = 2 and trt = 2 then x = x + &s2t2;
        output;
      end;
    end;
  end;
run;

%mend simulate;

*****;

%include "H:\dissertn\sas_macros\m_summary.sas";
%include "H:\dissertn\sas_macros\m_azzalini.sas";
%include "H:\dissertn\sas_macros\m_ciminera.sas";
%include "H:\dissertn\sas_macros\m_gailsimn.sas";
%include "H:\dissertn\sas_macros\m_recap.sas";
%include "H:\dissertn\sas_macros\m_saveperm.sas";

*****;
*****;

%macro do_it(seed, pattern, s1t1, s1t2, s2t1, s2t2, direct);

```

```

proc datasets library = WORK memtype = DATA KILL;
run;

data runtime;
  start = datetime();
  put 'START TIME = ' start datetime18.;
run;

*The following provides the sample size based on the DIRECT parameter;
%include "H:\dissertn\sas_macros\s2_sample_size.sas";

%put PATTERN = &pattern DATASET = &direct;

%do nsim = 1 %to 2500;

  %simulate;

  %summary(multi,x);
  %azzalini(bysite1);
  %ciminera(bysite1);
  %gailsimn(bysite1);

  %let nsim2 = &nsim;

%end;

options pageno = 1;
title "Interaction Pattern no. &pattern &direct Number of simulations is &nsim2";
title2 "Site 1 Trt 1 is &s1t1 Site 2 Trt 1 is &s2t1";
title3 "Site 1 Trt 2 is &s1t2 Site 2 Trt 2 is &s2t2";
title4 "Site 1 Sample size is &s1n Site 2 Sample Size is &s2n";
title5 "Seed is &seed";

%recap;

%saveperm;

data runtime;
  set runtime;
  if _N_ = 1 then end = datetime();
  runsec = (end - start);
  runmin = (end - start)/60;
  runhrs = (end - start)/(60*60);
  put 'RUNTIME = ' runsec 'seconds';

```

```

    put 'RUNTIME = ' runmin 'minutes';
    put 'RUNTIME = ' runhrs 'hours';
run;

%mend do_it;

*****;
*Note: Macro uses: Seed, pattern number, s1t1, s1t2, s2t1, s2t2 direct;
*Statements below have final randomization number;
*To change sample sizes remember to change DIRECT parameter;

%macro big(direct);

%do_it(878945, 1, 0.0, 0.5, 0.0, 0.5, &direct);
%do_it(878945, 2, 0.0, 0.0, 0.0, 1.0, &direct);
%do_it(878945, 3, 0.0, -0.25, 0.0, 1.25, &direct);
%do_it(878945, 4, 0.0, -0.5, 0.0, 1.5, &direct);
%do_it(878945, 5, 0.0, -1.0, 0.0, 2.0, &direct);
%do_it(878945, 6, 0.0, 0.0, 1.0, 2.0, &direct);
%do_it(878945, 7, 0.5, 0.0, 0.0, 1.5, &direct);
%do_it(878945, 8, 0.0, 0.25, 0.0, 0.75, &direct);
%do_it(878945, 9, -1.0, -0.5, 1.0, 1.5, &direct);
%do_it(878945, 10, 0.0, 1.0, 1.0, 0.0, &direct);
%do_it(878945, 11, 0.0, 0.0, 0.0, 0.0, &direct);

%mend big;

*****;

%big(S2A);
%big(S2B);
%big(S2C);
%big(S2F);
%big(S2G);

```

```

*m_master3.sas      13mar00;

*Simulation and Test Program for 3 centers;
*This program simulates data and analyzes it using the test macros;

*options ls = 78 ps = 52 pageno = 1 mprint mlogic symbolgen notes;
*options ls=78 ps=52 pageno=1 nomprint nomlogic nosymbolgen nosource;
options ls = 78 ps = 52 pageno = 1 nomprint nomlogic nosymbolgen nonotes;

%macro simulate;
data multi (keep = site trt npat x);
  seed2 = &seed + &nsim;
  do site = 1 to 3;
    if site = 1 then enrolled = &s1n;
    else if site = 2 then enrolled = &s2n;
    else if site = 3 then enrolled = &s3n;
    do trt = 1 to 2;
      do npat = 1 to enrolled;
        call rannor(seed2, x);
        if site = 1 and trt = 1 then x = x + &s1t1;
        if site = 1 and trt = 2 then x = x + &s1t2;
        if site = 2 and trt = 1 then x = x + &s2t1;
        if site = 2 and trt = 2 then x = x + &s2t2;
        if site = 3 and trt = 1 then x = x + &s3t1;
        if site = 3 and trt = 2 then x = x + &s3t2;
        output;
      end;
    end;
  end;
run;

%mend simulate;

*****;

%include "H:\dissertn\sas_macros\m_summary.sas";
%include "H:\dissertn\sas_macros\m_azzalini.sas";
%include "H:\dissertn\sas_macros\m_ciminera.sas";
%include "H:\dissertn\sas_macros\m_gailsimn.sas";
%include "H:\dissertn\sas_macros\m_recap.sas";
%include "H:\dissertn\sas_macros\m_saveperm.sas";

*****;
*****;

```

```

%macro do_it(seed, pattern,
             slt1, slt2, s2t1, s2t2, s3t1, s3t2, direct);

proc datasets library = WORK memtype = DATA KILL;
run;

data runtime;
  start = datetime();
  put 'START TIME = ' start datetime18.;
run;

*The following provides the sample size based on the DIRECT parameter;
%include "H:\dissertn\sas_macros\s3_sample_size.sas";

%put PATTERN = &pattern DATASET = &direct;

%do nsim = 1 %to 2500;

  %simulate;

  %summary(multi,x);
  %azzalini(bysite1);
  %ciminera(bysite1);
  %gailsimn(bysite1);

  %let nsim2 =&nsim;

%end;

options pageno = 1;
title "Interaction Pattern no. &pattern &direct Number of simulations is &nsim2";
title2 "Site 1 Trt 1 is &slt1 Site 2 Trt 1 is &s2t1 Site 3 Trt 1 is &s3t1";
title3 "Site 1 Trt 2 is &slt2 Site 2 Trt 2 is &s2t2 Site 3 Trt 2 is &s3t2";
title4 "Site 1 Sample size is &sln Site 2 Sample Size is &s2n Site 3 Sample Size
       is &s3n";
title5 "Seed is &seed";

%recap;

%saveperm;

data runtime;
  set runtime;

```

```

    if _N_ = 1 then end = datetime();
    runsec = (end - start);
    runmin = (end - start)/60;
    runhrs = (end - start)/(60*60);
    put 'RUNTIME = ' runsec 'seconds';
    put 'RUNTIME = ' runmin 'minutes';
    put 'RUNTIME = ' runhrs 'hours';
run;

%mend do_it;

*****;

*Note: Macro uses: Seed, pattern number;
*      slt1, slt2, s2t1, s2t2 s3t1 s3t2 direct;
*Statements below have final randomization number;
*To change sample sizes remember to change DIRECT parameter;

%macro big(direct);

%do_it(878945, 1, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, &direct);
%do_it(878945, 2, 0.0, 0.2, 0.0, 0.2, 0.0, 1.1, &direct);
%do_it(878945, 3, 0.0, 0.0, 0.0, 0.75, 0.0, 0.75, &direct);
%do_it(878945, 4, 0.0, 0.0, 0.0, 0.5, 0.0, 1.0, &direct);
%do_it(878945, 5, 0.0, 0.0, 0.0, 0.0, 0.0, 1.5, &direct);
%do_it(878945, 6, 0.0, -0.5, 0.0, 1.0, 0.0, 1.0, &direct);
%do_it(878945, 7, 0.0, -0.5, 0.0, 0.5, 0.0, 1.5, &direct);
%do_it(878945, 8, 0.0, -0.5, 0.0, 0.0, 0.0, 2.0, &direct);
%do_it(878945, 9, 0.0, -0.25, 0.0, -0.25, 0.0, 2.0, &direct);
%do_it(878945, 10, 0.0, -0.5, 0.0, 0.0, 0.0, 0.5, &direct);
%do_it(878945, 11, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, &direct);

%mend big;

*****;
*****;

%big(S3A);
%big(S3H);
%big(S3I);
%big(S3J);
%big(S3K);
%big(S3L);
%big(S3M);

```



```

*m_summary.sas                                12may00;

*This macro calculates treatment difference means, variances and df;
*for each site and other statistics needed for the test macros;

%macro summary(dataset,variable);

*****;
*The global macro variables defined below are used in other macros;

%global V_mean ov_N N_trts N_sites MSE3;

*****;
*****;
*This section uses Proc Mixed to test the Treatment by Center Interaction;
*using Analysis of Variance with the Type III model.
*It also provides the MSE (MSE3) and creates a macro variable;
*used later to create the variance of the mean based on the MSE;

*title "Analysis of variance using Proc Mixed";
*title2 "includes test for Treatment by Center interaction";

proc mixed data = &dataset noclprint noitprint;
  class site trt;
  model &variable = site trt site*trt;
  lsmeans trt site*trt / pdiff;
  make 'CovParms' out = var3new noprint;
  make 'Tests' out = aov3new noprint;
  make 'LSMeans' out = lsm3new noprint;
  make 'Diffs' out = dif3new noprint;
  make 'Fitting' out = fit3 noprint; *suppresses model fitting table;
run;

data var3new;
  set var3new (keep = est);
  call symput ('MSE3', EST);
  rename EST = T3_VAR;
run;

data aov3new (keep = T3_SITEF T3_TRT_F T3_S_T_F);
  retain T3_SITEF T3_TRT_F T3_S_T_F;
  set aov3new;
  if source = "SITE" then T3_SITEF = P_F;
  if source = "TRT" then T3_TRT_F = P_F;

```

```

if source = "SITE*TRT" then T3_S_T_F = P_F;
if _N_ = 3 then output;
run;

*****;
*****;

*This section uses Proc Mixed to test the Treatment difference;
*using Analysis of Variance with a Type II model.

*title "Analysis of variance using Proc Mixed";

proc mixed data = &dataset noclprint noitprint;
  class site trt;
  model &variable = site trt;
  * lsmeans trt / pdiff;
  make 'CovParms' out = var2new noprint;
  make 'Tests' out = aov2new noprint;
  * make 'Tests' out = aov2new;
  * make 'LSMeans' out = lsm2new noprint;
  * make 'Diffs' out = dif2new noprint;
  make 'Fitting' out = fit2 noprint; *suppresses model fitting table;
run;

data aov2new (keep = T2_SITEF T2_TRT_F);
  retain T2_SITEF T2_TRT_F;
  set aov2new;
  if source = "SITE" then T2_SITEF = P_F;
  if source = "TRT" then T2_TRT_F = P_F;
  if _N_ = 2 then output;
run;

*This section combines the summary output and creates the;
*summary dataset;

data var2new;
  set var2new (keep = EST);
  rename EST = T2_VAR;
run;

*****;
*****;

*This section calculates the overall sample size and
*creates a macro variable used in the Ciminera et al test;

```

```

proc sort data = &dataset;
  by site trt;
run;
proc univariate data = &dataset noprint;
  var &variable;
  output out = overall n = ov_N;
run;
data _null_;
  set overall;
  call symput ('ov_N', ov_N);
run;
*****;
*****;
*This section calculates for each treatment for each site;
*the mean, the number of observations and the variance;
*This can accommodate up to 5 treatments;
* and creates one record per site;
*There is no intrinsic limit to the number of sites;
*A macro variable for the pooled variance of the means across;
*all treatments is also created for use in the Azzalini macro;
*using an extension of the method used to calculate the pooled ;
*variance for a t-test (harmonic mean) (see notes of 13aug99);
*The variance of the means at each site is also calculated (var_mean);

proc summary data = &dataset;
  by site trt;
  var &variable;
  output out = bysite1 (keep = site trt mean var n)
    mean = mean var = var n = n;
run;

data bysite1;
  set bysite1;
  var_mean = &MSE3 / n;
run;

proc summary data = bysite1;
  var var_mean;
  output out = var_two
    sum = sum
    n = n;
run;

```

```

data _null_;
  set var_two;
  V_mean = sum / n;
  call symput ('V_mean', V_mean);
run;
*****;
*****;
*This section creates a single record for each site containing for each;
*treatment the mean, the number of observations and the variance;
*A macro variable for the number of treatments is also created;
*This can accommodate up to 5 treatments;
*There is no intrinsic limit to the number of sites;
*The means are printed and plotted;

*****;
*This section accommodates only two treatments;
data bysite1 (keep = site trt num_trts trt01-trt02 mean01-mean02 var_mean
               var01-var02 n01-n02);
  set bysite1;
  by site trt;
  array trts{*} $ 12 trt01-trt02;
  array means{*} mean01-mean02;
  array vars{*} var01-var02;
  array ns{*} n01-n02;
  if first.site then do;
    num_trts = 0;
    do in = 1 to 2;
      trts{in} = "";
      means{in} = .;
      vars{in} = .;
      ns{in} = .;
    end;
  end;
  retain num_trts trt01-trt02
         mean01-mean02
         var01-var02
         n01-n02;
  num_trts = num_trts + 1;
  trts{num_trts} = trt;
  means{num_trts} = mean;
  vars{num_trts} = var;
  ns{num_trts} = n;
  if last.site then output;
run;

```

```

data _null_;
  set bysite1;
  if _N_ = 1;
    call symput ('N_trts',compress(num_trts));
run;

proc summary data = bysite1;
  by trt;
  var mean01 - mean0&N_trts;
  output out = meannew (keep = mean01 - mean0&N_trts)
    mean = mean01 - mean0&N_trts;
run;

*proc append base = allmean data = meannew;
*run;

*****;
*****;

*This section calculates for each pair of treatments for each site;
*the mean difference, the degrees of freedom and the pooled variance;
*of the difference;
*also the variance of the difference using the MSE (MSE3);
*This can accommodate up to 5 treatments;
*There is no intrinsic limit to the number of sites;

data bysite1 (keep = site mean01 mean02 var_mean
                  dif0102
                  df0102
                  vsit0102
                  MSEv0102);

  set bysite1;
  by site;
  array difs{2,2}
    dif0101 - dif0102
    dif0201 - dif0202;
  array dfs{2,2}
    df0101 - df0102
    df0201 - df0202;
  array vsits{2,2}
    vsit0101 - vsit0102
    vsit0201 - vsit0202;

```

```

array MSEvsits{2,2}
    MSEv0101 - MSEv0102
    MSEv0201 - MSEv0202;
array means{2} mean01-mean02;
array vars{2} var01-var02;
array ns{2} n01-n02;
do iarr = 1 to (&N_trts - 1);
    do jarr = (iarr + 1) to &N_trts;
        difs{iarr,jarr} = means{iarr} - means{jarr};
        n_obs = ns{iarr} + ns{jarr};
        dfs{iarr,jarr} = n_obs - 2;
        f = 1 / ((1/ns{iarr}) + (1/ns{jarr}));
        v_pld = (((ns{iarr}-1)*vars{iarr}) + ((ns{jarr}-1)*vars{jarr}))
            / (n_obs-2);
        vsits{iarr,jarr} = v_pld/f;
        MSEvsits{iarr,jarr} = &MSE3/f;
    end;
end;
run;
*****;
*****;
*This section counts the number of sites and creates a macro variable;

proc summary data = bysite1;
    var site;
    output out = site_ct (keep = n1)
        n = n1;
run;

data _null_;
    set site_ct;
    call symput ('N_sites',compress(n1));
run;
*****;
*****;
*This section creates the dataset that contains the relevant differences;
*between of treatments within centers and the test results and merges them;
*with the ANOVA test results for the factors and the interaction;

%if &N_sites = 2 %then %do;

    data dif3new (keep = LSMD01 LSMD02 LSMP01 LSMP02);
        retain LSMD01 LSMD02 LSMP01 LSMP02;
        set dif3new;

```

```

if _effect_ = "SITE*TRT" and SITE="1" and TRT="1" and _SITE="1" and
  _TRT="2"
  then LSMD01 = _diff_;
if _effect_ = "SITE*TRT" and SITE="2" and TRT="1" and _SITE="2" and
  _TRT="2"
  then LSMD02 = _diff_;
if _effect_ = "SITE*TRT" and SITE="1" and TRT="1" and _SITE="1" and
  _TRT="2"
  then LSMP01 = _pt_;
if _effect_ = "SITE*TRT" and SITE="2" and TRT="1" and _SITE="2" and
  _TRT="2"
  then LSMP02 = _pt_;
if _N_ = 7 then output;
run;

%end;

%else %if &N_sites = 3 %then %do;

data dif3new (keep = LSMD01 LSMD02 LSMD03 LSMP01 LSMP02 LSMP03);
  retain LSMD01 LSMD02 LSMD03 LSMP01 LSMP02 LSMP03;
  set dif3new;
  if _effect_ = "SITE*TRT" and SITE="1" and TRT="1" and _SITE="1" and
    _TRT="2"
    then LSMD01 = _diff_;
  if _effect_ = "SITE*TRT" and SITE="2" and TRT="1" and _SITE="2" and
    _TRT="2"
    then LSMD02 = _diff_;
  if _effect_ = "SITE*TRT" and SITE="3" and TRT="1" and _SITE="3" and
    _TRT="2"
    then LSMD03 = _diff_;
  if _effect_ = "SITE*TRT" and SITE="1" and TRT="1" and _SITE="1" and
    _TRT="2"
    then LSMP01 = _pt_;
  if _effect_ = "SITE*TRT" and SITE="2" and TRT="1" and _SITE="2" and
    _TRT="2"
    then LSMP02 = _pt_;
  if _effect_ = "SITE*TRT" and SITE="3" and TRT="1" and _SITE="3" and
    _TRT="2"
    then LSMP03 = _pt_;
  if _N_ = 16 then output;
run;

%end;

```

```

data aov3new;
  merge aov3new dif3new;
run;

*****;

*This section combines the summary output and creates the;
*summary dataset;

data newsum;
  merge meannew var3new aov3new var2new aov2new;
  nsim = &nsim;
  acv_mean = &v_mean;
run;

proc append base = allsum data = newsum;
run;
*****;

%mend summary;

```



```

*m_azzalini.sas    11may00;

*These macros implement the test presented in Azzalini and Cox;
*The input data set needs to be in the bysite1 format;
*These macros have been adapted from the paper by Azzalini and Cox;
*to accomodate unequal sample sizes using two different methods;
*The first macro uses an exact testing procedure;
*The second macro uses an approximate testing procedure with the harmonic mean;
*sample size used to estimate the average sample size;
*Each of these macros is called by the third (controlling) macro;
*****;
*****;
%macro azz_x(dataset);

*****;
*This section accommodates two treatments only;
*****;
*This section sets up a duplicate array of pairwise differences;

*Note: The differences are multiplied by  $\sqrt{((df+2)/2)}$ , which is equal to;
* the  $\sqrt{}$  of the sample size of each mean used to calculate the difference;

data bysite2(keep = site dif0102)
    bysite3(keep = site dup0102);

set &dataset;

by site;
array dfs{2,2}
    df0101-df0102
    df0201-df0202;
array difs{2,2}
    dif0101-dif0102
    dif0201-dif0202;
array dups{2,2}
    dup0101-dup0102
    dup0201-dup0202;
do idup = 1 to 2;
    do jdup = 1 to 2;
        difs{idup,jdup} = (sqrt((dfs{idup,jdup} + 2)/2)) * difs{idup,jdup};
        dups{idup,jdup} = difs{idup,jdup};
    end;
end;
run;

```

```
*****;
```

```
*This section calculates the estimate of  $t = t\_hat$ ;
```

```
data xt_calc (keep = nsim x_maxdif MSE3 az_xt p_az_xt);
  maxdif = 0;
  retain maxdif;
  do point1 = 1 to &N_sites;
    set bysite2 point = point1;
    *array %list&N_trts(dif);
    array difs{2,2}
      dif0101-dif0102
      dif0201-dif0202;
    do i1 = 1 to (&N_trts - 1);
      do j1 = (i1 + 1) to &N_trts;
        element1 = difs{i1,j1};
        do point2 = (point1 + 1) to &N_sites;
          set bysite3 point = point2;
          *array %list&N_trts(dup);
          array dups{2,2}
            dup0101-dup0102
            dup0201-dup0202;
          element2 = dups{i1,j1};
          if (element1 le 0 and element2 ge 0) or
            (element1 ge 0 and element2 le 0) then
            t_min = min(abs(element1),abs(element2));
            if t_min gt maxdif then maxdif = t_min;
          end;
        end;
      end;
    end;
  xt_hat = maxdif/sqrt(2*&MSE3);
  MSE3 = &MSE3;
  az_xt = xt_hat;
  m1 = &N_sites;
  m2 = &N_trts;
  s0 = exp(-.5*(m1*(m1-1))*(m2*(m2-1))*((probnorm(-az_xt))**2));
  p_az_xt = 1 - s0;
  nsim = &nsim;
  x_maxdif = maxdif;
  output;
  stop;
run;
```

```

*****;

%mend azz_x;

*****;
*****;

%macro azz_app(dataset);

*****;

*This section accommodates two treatments only;

*****;

*This section sets up a duplicate array of pairwise differences;

data bysite2(keep = site dif0102)
  bysite3(keep = site dup0102);

set &dataset;

by site;
array difs{2,2}
  dif0101-dif0102
  dif0201-dif0202;
array dups{2,2}
  dup0101-dup0102
  dup0201-dup0202;
do idup = 1 to 2;
  do jdup = 1 to 2;
    dups{idup,jdup} = difs{idup,jdup};
  end;
end;
run;

*****;

*This section calculates the estimate of  $t = t_{\hat{}}$ ;

```

```

data at_calc (keep = a_maxdif a_V_mean az_at p_az_at);
  maxdif = 0;
  retain maxdif;
  do point1 = 1 to &N_sites;
    set bysite2 point = point1;
    array difs{2,2}
      dif0101-dif0102
      dif0201-dif0202;
    do i1 = 1 to (&N_trts - 1);
      do j1 = (i1 + 1) to &N_trts;
        element1 = difs{i1,j1};
        do point2 = (point1 + 1) to &N_sites;
          set bysite3 point = point2;
          array dups{2,2}
            dup0101-dup0102
            dup0201-dup0202;
          element2 = dups{i1,j1};

          if (element1 le 0 and element2 ge 0) or
            (element1 ge 0 and element2 le 0) then
            t_min = min(abs(element1),abs(element2));
            if t_min gt maxdif then maxdif = t_min;
          end;
        end;
      end;
    end;
  at_hat = maxdif/sqrt(2*&V_mean);
  a_V_mean = &V_mean;
  az_at = at_hat;
  m1 = &N_sites;
  m2 = &N_trts;
  s0 = exp(-.5*(m1*(m1-1))*(m2*(m2-1))*((probnorm(-az_at))**2));
  p_az_at = 1 - s0;
  a_maxdif = maxdif;
  output;
  stop;
run;

*****

%mend azz_app;

*****
*****

```

```
%macro azzalini(dataset);  
  
  %azz_x(&dataset);  
  %azz_app(&dataset);  
  
  data both_azz;  
    merge xt_calc at _calc;  
  run;  
  
  proc append base = allazz data = both_azz;  
  run;  
  
%mend azzalini;
```

```

*m_ciminera.sas      10may00;

*These macros implement the test presented in Ciminera et al.;
*The first uses a pooled estimate of error based on internally estimated variances;
*as defined in Ciminera et al;
*The second uses the MSE from a Type III ANOVA model;
*Only the output from the second method is presented in the dissertation;
*Each of these macros is called by the third (controlling) macro;
*The input data needs to be in bysite1 format;

*****-
*****-

%macro cim_pld(dataset, vnum);

*****-
*This macro contains local macro variables as defined below;
%local s_pld ov_Med;

*****-

*This section calculates the overall variance across all sites;
*from the respective site variances and creates a macro variable;

data bysite(keep = site dif df ss);
  set &dataset;
  dif = dif&vnum;
  df = df&vnum;
  ss = vsit&vnum * df&vnum;
run;

proc summary data = bysite;
  var ss df;
  output out = pooled (keep = sumvar sumdf)
    sum = sumvar sumdf;
run;

data _null_;
  set pooled;
  s_pooled = sqrt(sumvar / sumdf);
  call symput ('s_pld', s_pooled);
run;

*****-

```

```

*This section calculates the median difference across all sites;
*and creates a macro variable;
data difdata (keep = dif);
  set bysite;
  by site;
  do i = 1 to df + 2 ;
    output;
  end;
run;

proc univariate data = difdata noprint;
  var dif;
  output out = medover median = ov_Med;
run;
data _null_;
  set medover;
  call symput ('ov_Med', ov_Med);
run;

*****;

*This section produces the pushback procedure as displayed on;
*page 1040 of Ciminera et al;
*note that v1 and v2 are degrees of freedom and the not the variable numbers;
*denoted by the macro variables v1 and v2;
data pushback;
  set bysite;
  do sitn = 1 to &N_sites;
    step2 = dif - &ov_Med;
    pooled_s = &s_pld;
    step4 = step2 / &s_pld;
  end;
run;

proc sort data = pushback;
  by step2;
run;

data pushback (drop = ss sitn);
  set pushback;
  order = _N_;
  v1 = order;
  v2 = &N_sites - order + 1;

```

```

if step4 le 0 then
  do;
    p1 = betainv(.1,v1,v2);
  end;
else if step4 gt 0 then
  do;
    p1 = betainv(.9,v1,v2);
  end;
t1 = tinv(p1,df);
step8 = step4 - t1;
if step4 ge 0 and step8 lt 0 then step8A = 0;
else if step4 le 0 and step8 gt 0 then step8A = 0;
else step8A = step8;
step9 = step8A * &s_pld;
step10 = step9 + &ov_Med;
run;

*****;

*This section counts the number of positive and negative differences;
*before and after pushback and appends the counts to the summary dataset;

data rawnew(keep = rawneg rawzero rawpos s_pldpld nsim);
set bysite end = last;
if _n_=1 then do;
  rawneg = 0;
  rawzero = 0;
  rawpos = 0;
end;
if dif lt 0 then rawneg + 1;
else if dif eq 0 then rawzero + 1;
else if dif gt 0 then rawpos + 1;
s_pldpld = &s_pld;
nsim = &nsim;
if last then output;
run;

data cimnew (keep = sumneg sumzero sumpos);
set pushback end=last;
if _n_=1 then do;
  sumneg = 0;
  sumzero = 0;
  sumpos = 0;
end;

```



```

if step10 lt 0 then sumneg + 1;
else if step10 eq 0 then sumzero + 1;
else if step10 gt 0 then sumpos + 1;
if last then output;
run;

data cimpld;
merge cimnew rawnew;
rename sumneg = cimpldneg
       sumzero = cimpldzer
       sumpos = cimpldpos;
run;

*****;
%mend cim_pld;

*****;
*****;

%macro cim_mse(dataset, vnum);

*****;
*This macro contains local macro variables as defined below;
%local s_pld ov_Med;

*****;

*This section calculates the overall variance across all sites;
*from the respective site variances and creates a macro variable;

data bysite(keep = site dif df ss);
set &dataset;
dif = dif&vnum;
df = df&vnum;
ss = MSEv&vnum * df&vnum;
run;

proc summary data = bysite;
var ss df;
output out = pooled (keep = sumvar sumdf)
       sum = sumvar sumdf;
run;

```

```

data _null_;
  set pooled;
  s_pooled = sqrt(sumvar / sumdf);
  call symput ('s_pld', s_pooled);
run;
*****;
*This section calculates the median difference across all sites;
*and creates a macro variable;
data difdata (keep = dif);
  set bysite;
  by site;
  do i = 1 to df + 2 ;
    output;
  end;
run;

proc univariate data = difdata noprint;
  var dif;
  output out = medover median = ov_Med;
run;
data _null_;
  set medover;
  call symput ('ov_Med', ov_Med);
run;

*****;

*This section produces the pushback procedure as displayed on;
*page 1040 of Ciminera et al;
*note that v1 and v2 are degrees of freedom and the not the variable numbers;
*denoted by the macro variables v1 and v2;
data pushback;
  set bysite;
  do sitn = 1 to &N_sites;
    step2 = dif - &ov_Med;
    pooled_s = &s_pld;
    step4 = step2 / &s_pld;
  end;
run;

proc sort data = pushback;
  by step2;
run;

```

```

data pushback (drop = ss sitn);
  set pushback;
  order = _N_;
  v1 = order;
  v2 = &N_sites - order + 1;
  if step4 le 0 then
    do;
      pl = betainv(.1,v1,v2);
    end;
  else if step4 gt 0 then
    do;
      pl = betainv(.9,v1,v2);
    end;
  t1 = tinv(pl,df);
  step8 = step4 - t1;
  if step4 ge 0 and step8 lt 0 then step8A = 0;
  else if step4 le 0 and step8 gt 0 then step8A = 0;
  else step8A = step8;
  step9 = step8A * &s_pld;
  step10 = step9 + &ov_Med;
run;

*****;
*This section counts the number of positive and negative differences;
*after pushback and appends the counts to the summary dataset;

data rawnew(keep = s_pldmse);
  set bysite end = last;
  s_pldmse = &s_pld;
  if last then output;
run;

data cimnew (keep = sumneg sumzero sumpos);
  set pushback end=last;
  if _n_=1 then do;
    sumneg = 0;
    sumzero = 0;
    sumpos = 0;
  end;
  if step10 lt 0 then sumneg + 1;
  else if step10 eq 0 then sumzero + 1;
  else if step10 gt 0 then sumpos + 1;
  if last then output;
run;

```

```

data cimmse;
  merge cimnew rawnew;
  rename sumneg = cmmseneg
        sumzero = cmmsezer
        sumpos = cmmsepos;
run;

*****

%mend cim_mse;

*****
*****
*This macro runs the macros above for each pairwise set of treatment differences;
%macro ciminera(dataset);
  %do icim = 1 %to (&N_trts - 1);
    %do jcim = (&icim + 1) %to &N_trts;
      %let v1 = 0&icim;
      %let v2 = 0&jcim;
      %let vnum = &v1&v2;

      %cim_pld(&dataset, &vnum);
      %cim_mse(&dataset, &vnum);

      data cimnew;
        merge cimpld cimmse;
      run;

      proc append base = acim&vnum data = cimnew;
      run;

      %end;
    %end;
  %mend ciminera;

*****

```

```

*m_gailsimn.sas    12may00;

*These macros implement the test presented in Gail and Simon;
*The first uses a pooled estimate of error based on internally estimated variances;
*as defined in Ciminera et al;
*The second uses the MSE from a Type III ANOVA model;
*Only the output from the second method is presented in the dissertation;
*Each of these macros is called by the third (controlling) macro;

*****;
*****;

%macro gs_pld(dataset, vnum);

*****;
*This macro contains the local macro variables defined below;
*%local d_bar;

*****;
*This section calculates the ratios of differences and;
*variances of differences that will be used subsequently;
*in the calculations of Eq. 6 and the calculation of the Qs (Eq. 3);

data bysitegs (keep = site dif vdif prenum6 preden6 preQ);
  set &dataset;
  dif = dif&vnum;
  vdif = vsit&vnum;
  prenum6 = dif / vdif;
  preden6 = 1 / vdif;
  preQ = (dif)**2 / (vdif);
run;

*****;

*This section calculates dbar (Eq. 6) and H (Eq. 5);
*and calculates the test of heterogeneity based on H;
proc summary data = bysitegs;
  var prenum6 preden6;
  output out = pooledgs (keep = sumnum sumden)
    sum = sumnum sumden;
run;

```

```

data _null_;
  set pooledgs;
  d_bar = sumnum / sumden;
  call symput ('d_bar', d_bar);
run;

data difdatag (keep = eq5_site vdif);
  set bysitegs;
  den_5 = (dif - &d_bar)**2;
  eq5_site = (den_5 / vdif);
run;

proc summary data = difdatag;
  var eq5_site;
  output out = eq5_test (keep = Hpld)
    sum = Hpld;
run;

data hetero;
  set eq5_test;
  df = &N_sites - 1;
  P_Hpld = 1 - probchi(Hpld, df);
run;

*****;

*This section calculates and outputs the Qminus and Qplus statistics;
*The minimum of the two is also calculated;
*P values are calculated using the formulas given in the appendix of Gail and Simon;
proc summary data = bysitegs;
  where dif gt 0;
  var preQ;
  output out = Qminus (keep = Qminus)
    sum = Qminus;
run;

proc summary data = bysitegs;
  where dif lt 0;
  var preQ;
  output out = Qplus (keep = Qplus)
    sum = Qplus;
run;

```

```

*****-;

data Qpld (keep = nsim Hpld P_Hpld Qmnspld Qplspld minQpld P_Qmspld
      P_Qpspld P_mnQpld N_groups);
merge hetero Qminus Qplus;
N_groups = &N_sites;
nsim = &nsim;

if QMinus = . then QMinus = 0;
if QPlus = . then QPlus = 0;
  if QMinus le QPlus then minQ = QMinus;
  else if QPlus lt QMinus then minQ = QPlus;

nmq = N_groups - 1;
P_mnQpld = 0;
do mqindex = 1 to nmq;
  bprob = probbnml(0.5, nmq, mqindex) - probbnml(0.5, nmq, (mqindex -1)) ;
  cprob = 1 - probchi(minQ, mqindex);
  sub = bprob*cprob;
  P_mnQpld = P_mnQpld + sub;
end;

nqm = N_groups;
P_Qmspld = 0;
do qmindex = 1 to nqm;
  bprob = probbnml(0.5, nqm, qmindex) - probbnml(0.5, nqm, (qmindex -1)) ;
  cprob = 1 - probchi(QMinus, qmindex);
  sub = bprob*cprob;
  P_Qmspld = P_Qmspld + sub;
end;

nqp = N_groups;
P_Qpspld = 0;
do qpindex = 1 to nqp;
  bprob = probbnml(0.5, nqp, qpindex) - probbnml(0.5, nqp, (qpindex -1)) ;
  cprob = 1 - probchi(QPlus, qpindex);
  sub = bprob*cprob;
  P_Qpspld = P_Qpspld + sub;
end;

output;

```

```

rename Qminus = Qmnspld
      Qplus = Qplspld
      minQ = minQpld;

run;

*****;

%mend gs_pld;

*****;
*****;

%macro gs_MSE(dataset, vnum);

*****;
*This macro contains the local macro variables defined below;
*%local d_bar;

*****;

*This section calculates the ratios of differences and;
*variances of differences that will be used subsequently;
*in the calculations of Eq. 6 and the calculation of the Qs (Eq. 3);

data bysitegs (keep = site dif vdif prenum6 preden6 preQ);
  set &dataset;
  dif = dif&vnum;
  vdif = MSEv&vnum;
  prenum6 = dif / vdif;
  preden6 = 1 / vdif;
  preQ = (dif)**2 / (vdif);
run;

*****;

*This section calculates dbar (Eq. 6) and H (Eq. 5);
*and calculates the test of heterogeneity based on H;
proc summary data = bysitegs;
  var prenum6 preden6;
  output out = pooledgs (keep = sumnum sumden)
        sum = sumnum sumden;
run;

```



```

data _null_;
  set pooledgs;
  d_bar = sumnum / sumden;
  call symput ('d_bar', d_bar);
run;

data difdatag (keep = eq5_site vdif);
  set bysitegs;
  den_5 = (dif - &d_bar)**2;
  eq5_site = (den_5 / vdif);
run;

proc summary data = difdatag;
  var eq5_site;
  output out = eq5_test (keep = Hmse)
    sum = Hmse;
run;

data hetero;
  set eq5_test;
  df = &N_sites - 1;
  P_Hmse = 1 - probchi(Hmse, df);
run;

*****;
*This section calculates and outputs the Qminus and Qplus statistics;
*The minimum of the two is also calculated;
*P values are calculated using the formulas given in the appendix of Gail and Simon;
proc summary data = bysitegs;
  where dif gt 0;
  var preQ;
  output out = Qminus (keep = Qminus)
    sum = Qminus;
run;

proc summary data = bysitegs;
  where dif lt 0;
  var preQ;
  output out = Qplus (keep = Qplus)
    sum = Qplus;
run;

*****;

```

```

data Qmse (keep = Hmse P_Hmse Qmnsmse Qplsmse minQmse P_Qmsmse
      P_Qpsmse P_mnQmse);
merge hetero Qminus Qplus;
N_groups = &N_sites;

if QMinus = . then QMinus = 0;
if QPlus = . then QPlus = 0;
if QMinus le QPlus then minQ = QMinus;
else if QPlus lt QMinus then minQ = QPlus;

nmq = N_groups - 1;
P_mnQmse = 0;
do mqindex = 1 to nmq;
  bprob = probbnml(0.5, nmq, mqindex) - probbnml(0.5, nmq, (mqindex -1)) ;
  cprob = 1 - probchi(minQ, mqindex);
  sub = bprob*cprob;
  P_mnQmse = P_mnQmse + sub;
end;

nqm = N_groups;
P_Qmsmse = 0;
do qmindex = 1 to nqm;
  bprob = probbnml(0.5, nqm, qmindex) - probbnml(0.5, nqm, (qmindex -1)) ;
  cprob = 1 - probchi(QMinus, qmindex);
  sub = bprob*cprob;
  P_Qmsmse = P_Qmsmse + sub;
end;

nqp = N_groups;
P_Qpsmse = 0;
do qpindex = 1 to nqp;
  bprob = probbnml(0.5, nqp, qpindex) - probbnml(0.5, nqp, (qpindex -1)) ;
  cprob = 1 - probchi(QPlus, qpindex);
  sub = bprob*cprob;
  P_Qpsmse = P_Qpsmse + sub;
end;

output;

```

```

rename Qminus = Qmnsmse
      Qplus = Qplsmse
      minQ = minQmse;

run;

*****-,

%mend gs_MSE;

*****-,
*****-,

*This macro runs the macros above for each pairwise set of treatment differences;
%macro gailsimn(dataset);
  %do igs = 1 %to (&N_trts - 1);
    %do jgs = (&igs + 1) %to &N_trts;
      %let v1 = 0&igs;
      %let v2 = 0&jgs;
      %let vnum = &v1&v2;

      %gs_pld(&dataset, &vnum);
      %gs_MSE(&dataset, &vnum);

      data Qall;
        merge Qpld Qmse;
      run;

      proc append base = ags&vnum data = Qall;
      run;

      %end;
    %end;
  %mend gailsimn;

*****-,

```

```

*m_recap.sas      12may00;

*Recap program;
*This program summarizes (recaps) the results of the simulations;
*and the test results from the analyses of the simulations;

%macro recap;

title7 'Summary of data and Analysis of Variance Results';

data allsum;
  merge allsum acim0102(keep = nsim s_pldpld s_pldmse);
  by nsim;
run;

data allsum;
  set allsum;
  label T3_var = " "
        T2_var = " ";
  cim_vpld = s_pldpld*s_pldpld;
  cim_vmse = s_pldmse*s_pldmse;
  drop s_pldpld s_pldmse;
  if T3_TRT_F le 0.05 then T3TRT_05 = 100;
  else if T3_TRT_F gt 0.05 then T3TRT_05 = 0;
  if T2_TRT_F le 0.05 then T2TRT_05 = 100;
  else if T2_TRT_F gt 0.05 then T2TRT_05 = 0;
  if T3_S_T_F le 0.10 then T3_ST_10 = 100;
  else if T3_S_T_F gt 0.10 then T3_ST_10 = 0;
run;

proc means data = allsum;
  output out = asum_mns;
run;

title7 'Summary of Results of tests from Azzalini and Cox';

data allazz;
  set allazz;
  if p_az_xt le 0.10 then az_xt_10 = 100;
  else if p_az_xt gt 0.10 then az_xt_10 = 0;
  if p_az_at le 0.10 then az_at_10 = 100;
  else if p_az_at gt 0.10 then az_at_10 = 0;
run;

```

```

proc means data = allazz;
  output out = aazz_mns;
run;

title7 'Summary of Results of test from Ciminera et al.';
proc freq data = acim0102;
  tables (cmpldneg cmmseneg rawneg)
    / out = c0102frq;
run;

title7 'Summary of Results of tests from Gail and Simon';

data ags0102;
  set ags0102;
  if P_Hpld le 0.10 then Hpld_10 = 100;
  else if P_Hpld gt 0.10 then Hpld_10 = 0;
  if P_Hmse le 0.10 then Hmse_10 = 100;
  else if P_Hmse gt 0.10 then Hmse_10 = 0;
  if P_MnQpld le 0.10 then MnQp_10 = 100;
  else if P_MnQpld gt 0.10 then MnQp_10 = 0;
  if P_Qmspld le 0.025 then Qmsp_025 = 100;
  else if P_Qmspld gt 0.025 then Qmsp_025 = 0;
  if P_Qpspld le 0.025 then Qpsp_025 = 100;
  else if P_Qpspld gt 0.025 then Qpsp_025 = 0;
  if P_MnQmse le 0.10 then MnQm_10 = 100;
  else if P_MnQmse gt 0.10 then MnQm_10 = 0;
  if P_Qmsmse le 0.025 then Qmsm_025 = 100;
  else if P_Qmsmse gt 0.025 then Qmsm_025 = 0;
  if P_Qpsmse le 0.025 then Qpsm_025 = 100;
  else if P_Qpsmse gt 0.025 then Qpsm_025 = 0;
run;

proc means data = ags0102;
  output out = g0102mns;
run;

title7 'Comparison of Results Between Methods';

data alltests (keep = nsim T3_ST_10 az_xt_10 az_at_10 Hpld_10 Hmse_10
  MnQp_10 Qmsp_025 Qpsp_025 MnQm_10 Qmsm_025 Qpsm_025
  cmpldneg cmmseneg rawneg);
  merge allsum allazz acim0102 ags0102;
  by nsim;
run;

```

```

proc freq data = alltests;
  tables T3_ST_10 * (Hp1d_10 Hmse_10);
  tables az_xt_10 * (az_at_10 MnQp_10 MnQm_10 cmpldneg cmmseneg rawneg);
  tables      az_at_10 * (MnQp_10 MnQm_10 cmpldneg cmmseneg rawneg);
  tables      MnQp_10 * (MnQm_10 cmpldneg cmmseneg rawneg);
  tables      MnQm_10 * (cmpldneg cmmseneg rawneg);
  tables      cmpldneg * (cmmseneg rawneg);
  tables      cmmseneg * rawneg ;
run;

%mend recap;

```

```

*m_saveperm.sas      1dec99;

*This macro saves the raw data files and the data summaries;
* and the printed ouput;
* as permanent files;

%macro saveperm;
  libname out "H:\DISSERTN\SAS DATASETS\&direct\P&PATTERN";

  *This section saves the raw data files;
  data out.allsum;
    set allsum;
  run;
  data out.allazz;
    set allazz;
  run;
  data out.acim0102;
    set acim0102;
  run;
  data out.ags0102;
    set ags0102;
  run;

  *This section saves the data summaries;

  data out.sum_asum;
    set asum_mns;
  run;
  data out.sum_azz;
    set aazz_mns;
  run;
  data out.sumc0102;
    set c0102frq;
  run;
  data out.sumg0102;
    set g0102mns;
  run;

  *This section saves the printed output;

  %include "H:\dissertn\sas_macros\m_listsave.sas";
  %listsave;

%mend saveperm;

```

```

*m_listsave.sas      12may00;

*This macro is called by saveperm to save the SAS output;
* as a SAS formatted listing file with an MS WORD extension;

*options mprint mlogic symbolgen notes;
options pageno = 1;

%macro listsave;
  filename listing "H:\dissertn\sas_output\&direct\P&PATTERN..doc";

  proc printto new print = listing;
  run;

  title7 'Summary of data and Analysis of Variance Results';
  proc means data = allsum;
  run;
  title7 'Summary of Results of tests from Azzalini and Cox';
  proc means data = allazz;
  run;

  title7 'Summary of Results of test from Ciminera et al.';
  proc freq data = acim0102;
    tables (cmpldneg cmmseneg rawneg);
  run;

  title7 'Summary of Results of tests from Gail and Simon';
  proc means data = ags0102;
  run;

  title7 'Comparison of Results Between Methods';
  proc freq data = alltests;
    tables T3_ST_10 * (Hpld_10 Hmse_10);
    tables az_xt_10 * (az_at_10 MnQp_10 MnQm_10 cmpldneg cmmseneg rawneg);
    tables      az_at_10 * (MnQp_10 MnQm_10 cmpldneg cmmseneg rawneg);
    tables      MnQp_10 * (MnQm_10 cmpldneg cmmseneg rawneg);
    tables      MnQm_10 * (cmpldneg cmmseneg rawneg);
    tables      cmpldneg * (cmmseneg rawneg);
    tables      cmmseneg * rawneg ;
  run;

  proc printto;
  run;
%mend listsave;

```



```
*s2_sample_size.sas      03may00;

*Sample sizes for 2 centers;
*This section provides sample size options;
* to be determined by the designated DIRECT parameter;

%global s1n s2n;

%macro sampsize;

%if &direct = S2A %then
%do;
  %let s1n = 32;
  %let s2n = 32;
%end;

%if &direct = S2B %then
%do;
  %let s1n = 43;
  %let s2n = 21;
%end;

%if &direct = S2C %then
%do;
  %let s1n = 21;
  %let s2n = 43;
%end;

%if &direct = S2D %then
%do;
  %let s1n = 48;
  %let s2n = 16;
%end;

%if &direct = S2E %then
%do;
  %let s1n = 16;
  %let s2n = 48;
%end;
```

```
%if &direct = S2F %then
%do;
  %let s1n = 54;
  %let s2n = 10;
%end;

%if &direct = S2G %then
%do;
  %let s1n = 10;
  %let s2n = 54;
%end;
%mend sampsize;

%sampsize;
```

```

*s3_sample_size.sas      16dec99;

*Sample sizes for 3 centers;
*This section provides sample size options;
* to be determined by the designated DIRECT parameter;

%global s1n s2n s3n;

%macro sampsize;

%if &direct = S3A %then
%do;
    %let s1n = 22;
    %let s2n = 21;
    %let s3n = 21;
%end;

%if &direct = S3B %then
%do;
    %let s1n = 32;
    %let s2n = 16;
    %let s3n = 16;
%end;

%if &direct = S3C %then
%do;
    %let s1n = 16;
    %let s2n = 32;
    %let s3n = 16;
%end;

%if &direct = S3D %then
%do;
    %let s1n = 16;
    %let s2n = 16;
    %let s3n = 32;
%end;

%if &direct = S3E %then
%do;
    %let s1n = 26;
    %let s2n = 26;
    %let s3n = 12;
%end;

```

```
%if &direct = S3F %then
```

```
%do;
```

```
  %let s1n = 26;
```

```
  %let s2n = 12;
```

```
  %let s3n = 26;
```

```
%end;
```

```
%if &direct = S3G %then
```

```
%do;
```

```
  %let s1n = 12;
```

```
  %let s2n = 26;
```

```
  %let s3n = 26;
```

```
%end;
```

```
%if &direct = S3H %then
```

```
%do;
```

```
  %let s1n = 39;
```

```
  %let s2n = 13;
```

```
  %let s3n = 12;
```

```
%end;
```

```
%if &direct = S3I %then
```

```
%do;
```

```
  %let s1n = 13;
```

```
  %let s2n = 39;
```

```
  %let s3n = 12;
```

```
%end;
```

```
%if &direct = S3J %then
```

```
%do;
```

```
  %let s1n = 13;
```

```
  %let s2n = 12;
```

```
  %let s3n = 39;
```

```
%end;
```

```
%if &direct = S3K %then
```

```
%do;
```

```
  %let s1n = 29;
```

```
  %let s2n = 29;
```

```
  %let s3n = 6;
```

```
%end;
```

```
%if &direct = S3L %then
```

```
%do;
```

```
  %let s1n = 29;
```

```
  %let s2n = 6;
```

```
  %let s3n = 29;
```

```
%end;
```

```
%if &direct = S3M %then
```

```
%do;
```

```
  %let s1n = 6;
```

```
  %let s2n = 29;
```

```
  %let s3n = 29;
```

```
%end;
```

```
%mend sampsize;
```

```
%sampsize;
```

```

*m_relate2.sas      22aug00;

*Relate program;
*This program summarizes (relates) the results of the simulations;
*and the test results from the analyses of the simulations;
*Specifically, it summarizes the relationships between positive results;

options mprint mlogic symbolgen notes;
*options nomprint nomlogic nosymbolgen nonotes;

*****;
%macro relate(pattern);

libname indata "H:\DISSERTN\SAS DATASETS\&direct\P&PATTERN";

data patID;
  length PATTERN $3;
  pattern = "P&PATTERN";
run;

*****;
*This section merges the output data sets created by the recap macro ;
*****;

data alltests (keep = nsim T3_ST_10 HPLD_10 HMSE_10 AZ_XT_10 AZ_AT_10
                  MNQP_10 MNQM_10 CMPLDNEG CMMSENEG RAWNEG);
  merge indata.allsum indata.allazz indata.acim0102 indata.ags0102;
  by nsim;
run;

*****;
*This section creates output data sets with cross tabulations of the ;
*results of the different methods ;
*This section calculates frequency of agreement of the ;
*results of the different methods ;
*****;

%macro agree22(var1, var2, dataset, out1, out2, outagree);

proc freq data = alltests;
  tables &var1 * &var2 / noprint out = &dataset;
run;

```

```

data &dataset (keep = &out1 &out2 &outagree);
  set &dataset end = last;
  retain P0_0 P0_100 P100_0 P100_100;
  if &var1 = 0 and &var2 = 0 then P0_0 = PERCENT;
  if &var1 = 0 and &var2 = 100 then P0_100 = PERCENT;
  if &var1 = 100 and &var2 = 0 then P100_0 = PERCENT;
  if &var1 = 100 and &var2 = 100 then P100_100 = PERCENT;

  if P0_0 = . then P0_0 = 0;
  if P0_100 = . then P0_100 = 0;
  if P100_0 = . then P100_0 = 0;
  if P100_100 = . then P100_100 = 0;

  &out1 = P100_0 + P100_100;
  &out2 = P0_100 + P100_100;
  &outagree = P0_0 + P100_100;
  if last then output;
run;

%mend agree22;

*****;

%macro agree23(var1, var2, dataset, out1, out2, outagree);

proc freq data = alltests;
  tables &var1 * &var2 / noprint out = &dataset;
run;

data &dataset (keep = &out1 &out2 &outagree);
  set &dataset end = last;
  retain P0_0 P0_1 P0_2 P100_0 P100_1 P100_2;
  if &var1 = 0 and &var2 = 0 then P0_0 = PERCENT;
  if &var1 = 0 and &var2 = 1 then P0_1 = PERCENT;
  if &var1 = 0 and &var2 = 2 then P0_2 = PERCENT;
  if &var1 = 100 and &var2 = 0 then P100_0 = PERCENT;
  if &var1 = 100 and &var2 = 1 then P100_1 = PERCENT;
  if &var1 = 100 and &var2 = 2 then P100_2 = PERCENT;

```

```

if P0_0 = . then P0_0 = 0;
if P0_1 = . then P0_1 = 0;
if P0_2 = . then P0_2 = 0;
if P100_0 = . then P100_0 = 0;
if P100_1 = . then P100_1 = 0;
if P100_2 = . then P100_2 = 0;

&out1 = P100_0 + P100_1 + P100_2;
&out2 = P100_1 + P0_1;
&outagree = P0_0 + P0_2 + P100_1;
if last then output;
run;

%mend agree23;

*****;
%macro agree33(var1, var2, dataset, out1, out2, outagree);

proc freq data = alltests;
  tables &var1 * &var2 / noprint out = &dataset;
run;

data &dataset (keep = &out1 &out2 &outagree);
  set &dataset end = last;
  retain P0_0 P0_1 P0_2 P1_0 P1_1 P1_2 P2_0 P2_1 P2_2;
  if &var1 = 0 and &var2 = 0 then P0_0 = PERCENT;
  if &var1 = 0 and &var2 = 1 then P0_1 = PERCENT;
  if &var1 = 0 and &var2 = 2 then P0_2 = PERCENT;
  if &var1 = 1 and &var2 = 0 then P1_0 = PERCENT;
  if &var1 = 1 and &var2 = 1 then P1_1 = PERCENT;
  if &var1 = 1 and &var2 = 2 then P1_2 = PERCENT;
  if &var1 = 2 and &var2 = 0 then P2_0 = PERCENT;
  if &var1 = 2 and &var2 = 1 then P2_1 = PERCENT;
  if &var1 = 2 and &var2 = 2 then P2_2 = PERCENT;

  if P0_0 = . then P0_0 = 0;
  if P0_1 = . then P0_1 = 0;
  if P0_2 = . then P0_2 = 0;
  if P1_0 = . then P1_0 = 0;
  if P1_1 = . then P1_1 = 0;
  if P1_2 = . then P1_2 = 0;
  if P2_0 = . then P2_0 = 0;
  if P2_1 = . then P2_1 = 0;
  if P2_2 = . then P2_2 = 0;

```



```

&out1   = P1_0 + P1_1 + P1_2;
&out2   = P2_1 + P1_1 + P0_1;
&outagree = P0_0 + P1_1 + P2_2;
if last then output;
run;

```

```
%mend agree33;
```

```
*****;
```

```

%agree22(T3_ST_10, HPLD_10, CT_GSH_P, T3_ST_10, HPLD_10,
         CT_GSH_P);
%agree22(T3_ST_10, HMSE_10, CT_GSH_M, T3_ST_10, HMSE_10,
         CT_GSH_M);
%agree22(HPLD_10, HMSE_10, GSHGSHPM, HPLD_10, HMSE_10,
         GSHGSHPM);
%agree22(AZ_XT_10, AZ_AT_10, AC_AC_XA, AZ_XT_10, AZ_AT_10,
         AC_AC_XA);
%agree22(AZ_XT_10, MNQP_10, AC_GS_XP, AZ_XT_10, MNQP_10,
         AC_GS_XP);
%agree22(AZ_XT_10, MNQM_10, AC_GS_XM, AZ_XT_10, MNQM_10,
         AC_GS_XM);
%agree22(AZ_AT_10, MNQP_10, AC_GS_AP, AZ_AT_10, MNQP_10,
         AC_GS_AP);
%agree22(AZ_AT_10, MNQM_10, AC_GS_AM, AZ_AT_10, MNQM_10,
         AC_GS_AM);
%agree23(AZ_XT_10, CMPLDNEG, AC_CM_XP, AZ_XT_10, CMPLD_SE,
         AC_CM_XP);
%agree23(AZ_XT_10, CMMSENEG, AC_CM_XM, AZ_XT_10, CMMSE_SE,
         AC_CM_XM);
%agree23(AZ_AT_10, CMPLDNEG, AC_CM_AP, AZ_AT_10, CMPLD_SE,
         AC_CM_AP);
%agree23(AZ_AT_10, CMMSENEG, AC_CM_AM, AZ_AT_10, CMMSE_SE,
         AC_CM_AM);
%agree23(AZ_XT_10, RAWNEG, AC_RAW_X, AZ_XT_10, RAW_SE,
         AC_RAW_X);
%agree23(AZ_AT_10, RAWNEG, AC_RAW_A, AZ_AT_10, RAW_SE,
         AC_RAW_A);
%agree22(MNQP_10, MNQM_10, GS_GS_PM, MNQP_10, MNQM_10,
         GS_GS_PM);
%agree23(MNQP_10, CMPLDNEG, GS_CM_PP, MNQP_10, CMPLD_SE,
         GS_CM_PP);

```

```

%agree23(MNQP_10, CMMSENEG, GS_CM_PM, MNQP_10, CMMSE_SE,
          GS_CM_PM);
%agree23(MNQM_10, CMPLDNEG, GS_CM_MP, MNQM_10, CMPLD_SE,
          GS_CM_MP);
%agree23(MNQM_10, CMMSENEG, GS_CM_MM, MNQM_10, CMMSE_SE,
          GS_CM_MM);
%agree23(MNQP_10, RAWNEG, GS_RAW_P, MNQP_10, RAW_SE,
          GS_RAW_P);
%agree23(MNQM_10, RAWNEG, GS_RAW_M, MNQM_10, RAW_SE,
          GS_RAW_M);
%agree33(CMPLDNEG, CMMSENEG, CM_CM_PM, CMPLD_SE, CMMSE_SE,
          CM_CM_PM);
%agree33(CMPLDNEG, RAWNEG, CM_RAW_P, CMPLD_SE, RAW_SE,
          CM_RAW_P);
%agree33(CMMSENEG, RAWNEG, CM_RAW_M, CMMSE_SE, RAW_SE,
          CM_RAW_M);

*****.

data rel_new;
merge patID CT_GSH_P CT_GSH_M GSHGSHPM
          AC_AC_XA AC_GS_XP AC_GS_XM AC_GS_AP AC_GS_AM
            AC_CM_XP AC_CM_XM AC_CM_AP AC_CM_AM AC_RAW_X
            AC_RAW_A
          GS_GS_PM GS_CM_PP GS_CM_PM GS_CM_MP GS_CM_MM
            GS_RAW_P GS_RAW_M
          CM_CM_PM CM_RAW_P CM_RAW_M;
FORMAT T3_ST_10 HPLD_10 HMSE_10 CT_GSH_P CT_GSH_M GSHGSHPM
          AZ_XT_10 AZ_AT_10 MNQP_10 MNQM_10 CMPLD_SE CMMSE_SE
            RAW_SE
          AC_AC_XA AC_GS_XP AC_GS_XM AC_GS_AP AC_GS_AM
            AC_CM_XP AC_CM_XM AC_CM_AP AC_CM_AM AC_RAW_X
            AC_RAW_A
          GS_GS_PM GS_CM_PP GS_CM_PM GS_CM_MP GS_CM_MM
            GS_RAW_P GS_RAW_M
          CM_CM_PM CM_RAW_P CM_RAW_M
3.;

proc append base = a_relate data = rel_new;
run;

*****.
*This section captures summary statistics for the current pattern and;
*writes them into a dataset with summaries of all patterns for group ;

```

```

*****;
data rstats (keep = mean_t01 mean_t02 ac_var cim_vpld cim_vmse t3_var t2_var
               t3trt_05 t2trt_05 qmsp_025 qpssp_025 qmsm_025 qpsm_025);
merge indata.sum_asum (keep = _stat_ mean01 mean02 acv_mean cim_vpld
               cim_vmse
               t3_var t2_var t3trt_05 t2trt_05)
      indata.sumg0102 (keep = _stat_ qmsp_025 qpssp_025 qmsm_025 qpsm_025);
where _STAT_ = "MEAN";
AC_VAR = 2*acv_mean;
MEAN_T01 = mean01;
MEAN_T02 = mean02;
format mean_t01 mean_t02 ac_var cim_vpld cim_vmse t3_var t2_var 6.3;
format t3trt_05 t2trt_05 qmsp_025 qpssp_025 qmsm_025 qpsm_025 3.;
run;

data rstats;
merge patID rstats;
run;

proc append base = a_rstats data = rstats;
run;

%mend relate;
*****;
*****;
%macro do_all(direct);

proc datasets library = WORK memtype = DATA KILL;
run;

title7 "Comparison of Results Between Methods &DIRECT";
title8 "Patterns with Duplicate Results (9,6,7) omitted";

%relate(1);
* %relate(9);
%relate(8);
%relate(2);
* %relate(6);
%relate(3);
%relate(4);
* %relate(7);
%relate(5);
%relate(11);
%relate(10);

```

```

options pageno = 1;
*****;
filename outdoc "H:\dissertn\sas_output\&direct\relate2test.doc";

proc printto new print = outdoc;
run;

proc print data = a_rstats noobs;
  var pattern mean_t01 mean_t02 ac_var cim_vpld cim_vmse t3_var t2_var
    pattern t3trt_05 t2trt_05 qmsp_025 qpsp_025 qmsm_025 qpsm_025;
run;

proc print data = a_relate noobs;
  var PATTERN T3_ST_10 HPLD_10 HMSE_10 CT_GSH_P CT_GSH_M
    GSHGSHPM;
run;

proc print data = a_relate noobs;
  var PATTERN AZ_XT_10 AZ_AT_10 MNQP_10 MNQM_10 CMPLD_SE
    CMMSE_SE RAW_SE;
run;

proc print data = a_relate noobs;
  var PATTERN
    AC_AC_XA AC_GS_XP AC_GS_XM AC_GS_AP AC_GS_AM
    AC_CM_XP AC_CM_XM AC_CM_AP AC_CM_AM
    GS_GS_PM GS_CM_PP GS_CM_PM GS_CM_MP GS_CM_MM
    CM_CM_PM;
run;

proc print data = a_relate noobs;
  var PATTERN AC_RAW_X AC_RAW_A GS_RAW_P GS_RAW_M
    CM_RAW_P CM_RAW_M;
run;

proc printto;
run;
*****
proc print data = a_rstats noobs;
  var pattern mean_t01 mean_t02 ac_var cim_vpld cim_vmse t3_var t2_var
    pattern t3trt_05 t2trt_05 qmsp_025 qpsp_025 qmsm_025 qpsm_025;
run;

```

```

proc print data = a_relate noobs;
  var PATTERN T3_ST_10 HPLD_10 HMSE_10 CT_GSH_P CT_GSH_M
    GSHGSHPM;
run;

proc print data = a_relate noobs;
  var PATTERN AZ_XT_10 AZ_AT_10 MNQP_10 MNQM_10 CMPLD_SE
    CMMSE_SE RAW_SE;
run;

proc print data = a_relate noobs;
  var PATTERN
    AC_AC_XA AC_GS_XP AC_GS_XM AC_GS_AP AC_GS_AM
    AC_CM_XP AC_CM_XM AC_CM_AP AC_CM_AM
    GS_GS_PM GS_CM_PP GS_CM_PM GS_CM_MP GS_CM_MM
    CM_CM_PM;
run;

proc print data = a_relate noobs;
  var PATTERN AC_RAW_X AC_RAW_A GS_RAW_P GS_RAW_M
    CM_RAW_P CM_RAW_M;
run;

%mend do_all;
*****-,

%do_all(S2A);
%do_all(S2B);
%do_all(S2C);
%do_all(S2F);
%do_all(S2G);

```

```

*m_relate3.sas      23dec00;

*Relate program;
*This program summarizes (relates) the results of the simulations;
*and the test results from the analyses of the simulations;
*Specifically, it summarizes the relationships between positive results;

*options mprint mlogic symbolgen notes;
options nomprint nomlogic nosymbolgen nonotes;

*****;
%macro relate(pattern);

libname indata "H:\DISSERTN\SAS DATASETS\&direct\P&PATTERN";

data patID;
  length PATTERN $3;
  pattern = "P&PATTERN";
run;

*****;
*This section merges the output data sets created by the recap macro ;
*****;

data alltests (keep = nsim T3_ST_10 HPLD_10 HMSE_10 AZ_XT_10 AZ_AT_10
                  MNQP_10 MNQM_10 CMPLDNEG CMMSeneg RAWNEG);
  merge indata.allsum indata.allazz indata.acim0102 indata.ags0102;
  by nsim;
run;

*****;
*This section creates output data sets with cross tabulations of the ;
*results of the different methods ;
*This section calculates frequency of agreement of the ;
*results of the different methods ;
*****;

%macro agree22(var1, var2, dataset, out1, out2, outagree);

proc freq data = alltests;
  tables &var1 * &var2 / noprint out = &dataset;
run;

```

```

data &dataset (keep = &out1 &out2 &outagree);
  set &dataset end = last;
  retain P0_0 P0_100 P100_0 P100_100;
  if &var1 = 0 and &var2 = 0 then P0_0 = PERCENT;
  if &var1 = 0 and &var2 = 100 then P0_100 = PERCENT;
  if &var1 = 100 and &var2 = 0 then P100_0 = PERCENT;
  if &var1 = 100 and &var2 = 100 then P100_100 = PERCENT;

  if P0_0 = . then P0_0 = 0;
  if P0_100 = . then P0_100 = 0;
  if P100_0 = . then P100_0 = 0;
  if P100_100 = . then P100_100 = 0;

  &out1 = P100_0 + P100_100;
  &out2 = P0_100 + P100_100;
  &outagree = P0_0 + P100_100;
  if last then output;
run;

%mend agree22;

*****;

%macro agree24(var1, var2, dataset, out1, out2, outagree);

proc freq data = alltests;
  tables &var1 * &var2 / noprint out = &dataset;
run;

data &dataset (keep = &out1 &out2 &outagree);
  set &dataset end = last;
  retain P0_0 P0_1 P0_2 P0_3 P100_0 P100_1 P100_2 P100_3;
  if &var1 = 0 and &var2 = 0 then P0_0 = PERCENT;
  if &var1 = 0 and &var2 = 1 then P0_1 = PERCENT;
  if &var1 = 0 and &var2 = 2 then P0_2 = PERCENT;
  if &var1 = 0 and &var2 = 3 then P0_3 = PERCENT;
  if &var1 = 100 and &var2 = 0 then P100_0 = PERCENT;
  if &var1 = 100 and &var2 = 1 then P100_1 = PERCENT;
  if &var1 = 100 and &var2 = 2 then P100_2 = PERCENT;
  if &var1 = 100 and &var2 = 3 then P100_3 = PERCENT;

```

```

if P0_0 = . then P0_0 = 0;
if P0_1 = . then P0_1 = 0;
if P0_2 = . then P0_2 = 0;
if P0_3 = . then P0_3 = 0;
if P100_0 = . then P100_0 = 0;
if P100_1 = . then P100_1 = 0;
if P100_2 = . then P100_2 = 0;
if P100_3 = . then P100_3 = 0;

&out1 = P100_0 + P100_1 + P100_2 + P100_3; *can be commented out later;
&out2 = P100_1 + P0_1 + P100_2 + P0_2;
&outagree = P0_0 + P0_3 + P100_1 + P100_2;
if last then output;
run;

%mend agree24;

*****-,
%macro agree44(var1, var2, dataset, out1, out2, outagree);

proc freq data = alltests;
  tables &var1 * &var2 / noprint out = &dataset;
run;

data &dataset (keep = &out1 &out2 &outagree);
  set &dataset end = last;
  retain P0_0 P0_1 P0_2 P0_3 P1_0 P1_1 P1_2 P1_3
         P2_0 P2_1 P2_2 P2_3 P3_0 P3_1 P3_2 P3_3;
  if &var1 = 0 and &var2 = 0 then P0_0 = PERCENT;
  if &var1 = 0 and &var2 = 1 then P0_1 = PERCENT;
  if &var1 = 0 and &var2 = 2 then P0_2 = PERCENT;
  if &var1 = 0 and &var2 = 3 then P0_3 = PERCENT;
  if &var1 = 1 and &var2 = 0 then P1_0 = PERCENT;
  if &var1 = 1 and &var2 = 1 then P1_1 = PERCENT;
  if &var1 = 1 and &var2 = 2 then P1_2 = PERCENT;
  if &var1 = 1 and &var2 = 3 then P1_3 = PERCENT;
  if &var1 = 2 and &var2 = 0 then P2_0 = PERCENT;
  if &var1 = 2 and &var2 = 1 then P2_1 = PERCENT;
  if &var1 = 2 and &var2 = 2 then P2_2 = PERCENT;
  if &var1 = 2 and &var2 = 3 then P2_3 = PERCENT;
  if &var1 = 3 and &var2 = 0 then P3_0 = PERCENT;
  if &var1 = 3 and &var2 = 1 then P3_1 = PERCENT;
  if &var1 = 3 and &var2 = 2 then P3_2 = PERCENT;
  if &var1 = 3 and &var2 = 3 then P3_3 = PERCENT;

```



```

if P0_0 = . then P0_0 = 0;
if P0_1 = . then P0_1 = 0;
if P0_2 = . then P0_2 = 0;
if P0_3 = . then P0_3 = 0;
if P1_0 = . then P1_0 = 0;
if P1_1 = . then P1_1 = 0;
if P1_2 = . then P1_2 = 0;
if P1_3 = . then P1_3 = 0;
if P2_0 = . then P2_0 = 0;
if P2_1 = . then P2_1 = 0;
if P2_2 = . then P2_2 = 0;
if P2_3 = . then P2_3 = 0;
if P3_0 = . then P3_0 = 0;
if P3_1 = . then P3_1 = 0;
if P3_2 = . then P3_2 = 0;
if P3_3 = . then P3_3 = 0;

&out1   = P1_0 + P1_1 + P1_2 + P1_3 + P2_0 + P2_1 + P2_2 + P2_3;
&out2   = P3_1 + P2_1 + P1_1 + P0_1 + P3_2 + P2_2 + P1_2 + P0_2;
&outagree = P0_0 + P0_3 + P1_1 + P1_2 + P2_1 + P2_2 + P3_0 + P3_3;
  if last then output;
run;

%mend agree44;
*****;

%agree22(T3_ST_10, HPLD_10, CT_GSH_P, T3_ST_10, HPLD_10,
  CT_GSH_P);
%agree22(T3_ST_10, HMSE_10, CT_GSH_M, T3_ST_10, HMSE_10,
  CT_GSH_M);
%agree22(HPLD_10, HMSE_10, GSHGSHPM, HPLD_10, HMSE_10,
  GSHGSHPM);

%agree22(AZ_XT_10, AZ_AT_10, AC_AC_XA, AZ_XT_10, AZ_AT_10,
  AC_AC_XA);
%agree22(AZ_XT_10, MNQP_10, AC_GS_XP, AZ_XT_10, MNQP_10,
  AC_GS_XP);
%agree22(AZ_XT_10, MNQM_10, AC_GS_XM, AZ_XT_10, MNQM_10,
  AC_GS_XM);
%agree22(AZ_AT_10, MNQP_10, AC_GS_AP, AZ_AT_10, MNQP_10,
  AC_GS_AP);
%agree22(AZ_AT_10, MNQM_10, AC_GS_AM, AZ_AT_10, MNQM_10,
  AC_GS_AM);

```

```

%agree24(AZ_XT_10, CMPLDNEG, AC_CM_XP, AZ_XT_10, CMPLD_SE,
          AC_CM_XP);
%agree24(AZ_XT_10, CMMSSENEG, AC_CM_XM, AZ_XT_10, CMMSE_SE,
          AC_CM_XM);
%agree24(AZ_AT_10, CMPLDNEG, AC_CM_AP, AZ_AT_10, CMPLD_SE,
          AC_CM_AP);
%agree24(AZ_AT_10, CMMSSENEG, AC_CM_AM, AZ_AT_10, CMMSE_SE,
          AC_CM_AM);
%agree24(AZ_XT_10, RAWNEG, AC_RAW_X, AZ_XT_10, RAW_SE,
          AC_RAW_X);
%agree24(AZ_AT_10, RAWNEG, AC_RAW_A, AZ_AT_10, RAW_SE,
          AC_RAW_A);
%agree22(MNQP_10, MNQM_10, GS_GS_PM, MNQP_10, MNQM_10,
          GS_GS_PM);
%agree24(MNQP_10, CMPLDNEG, GS_CM_PP, MNQP_10, CMPLD_SE,
          GS_CM_PP);
%agree24(MNQP_10, CMMSSENEG, GS_CM_PM, MNQP_10, CMMSE_SE,
          GS_CM_PM);
%agree24(MNQM_10, CMPLDNEG, GS_CM_MP, MNQM_10, CMPLD_SE,
          GS_CM_MP);
%agree24(MNQM_10, CMMSSENEG, GS_CM_MM, MNQM_10, CMMSE_SE,
          GS_CM_MM);
%agree24(MNQP_10, RAWNEG, GS_RAW_P, MNQP_10, RAW_SE,
          GS_RAW_P);
%agree24(MNQM_10, RAWNEG, GS_RAW_M, MNQM_10, RAW_SE,
          GS_RAW_M);
%agree44(CMPLDNEG, CMMSSENEG, CM_CM_PM, CMPLD_SE, CMMSE_SE,
          CM_CM_PM);
%agree44(CMPLDNEG, RAWNEG, CM_RAW_P, CMPLD_SE, RAW_SE,
          CM_RAW_P);
%agree44(CMMSSENEG, RAWNEG, CM_RAW_M, CMMSE_SE, RAW_SE,
          CM_RAW_M);
*****;

data rel_new;
merge patID CT_GSH_P CT_GSH_M GSHGSHPM
          AC_AC_XA AC_GS_XP AC_GS_XM AC_GS_AP AC_GS_AM
            AC_CM_XP AC_CM_XM AC_CM_AP AC_CM_AM AC_RAW_X
            AC_RAW_A
          GS_GS_PM GS_CM_PP GS_CM_PM GS_CM_MP GS_CM_MM
            GS_RAW_P GS_RAW_M
          CM_CM_PM CM_RAW_P CM_RAW_M;

```

```

FORMAT T3_ST_10 HPLD_10 HMSE_10 CT_GSH_P CT_GSH_M GSHGSHPM
      AZ_XT_10 AZ_AT_10 MNQP_10 MNQM_10 CMPLD_SE CMMSE_SE
      RAW_SE
      AC_AC_XA AC_GS_XP AC_GS_XM AC_GS_AP AC_GS_AM
      AC_CM_XP AC_CM_XM AC_CM_AP AC_CM_AM AC_RAW_X
      AC_RAW_A
      GS_GS_PM GS_CM_PP GS_CM_PM GS_CM_MP GS_CM_MM
      GS_RAW_P GS_RAW_M
      CM_CM_PM CM_RAW_P CM_RAW_M
3.;

proc append base = a_relate data = rel_new;
run;

*****;
*This section captures summary statistics for the current pattern and;
*writes them into a dataset with summaries of all patterns for group ;
*****;

data rstats (keep = mean_t01 mean_t02 ac_var cim_vpld cim_vmse t3_var t2_var
               t3trt_05 t2trt_05 qmsp_025 qpssp_025 qmsm_025 qpasm_025);
  merge indata.sum_asum (keep = _stat_mean01 mean02 acv_mean cim_vpld
                           cim_vmse
                           t3_var t2_var t3trt_05 t2trt_05)
        indata.sumg0102 (keep = _stat_qmsp_025 qpssp_025 qmsm_025 qpasm_025);
  where _STAT_ = "MEAN";
  AC_VAR = 2*acv_mean;
  MEAN_T01 = mean01;
  MEAN_T02 = mean02;
  format mean_t01 mean_t02 ac_var cim_vpld cim_vmse t3_var t2_var 6.3;
  format t3trt_05 t2trt_05 qmsp_025 qpssp_025 qmsm_025 qpasm_025 3.;
run;

data rstats;
  merge patID rstats;
run;

proc append base = a_rstats data = rstats;
run;

%mend relate;
*****;
*****;

```

```

%macro do_all(direct);

proc datasets library = WORK memtype = DATA KILL;
run;

title7 "Comparison of Results Between Methods  &DIRECT";

%do ipat = 1 %to 11;
  %relate(&ipat);
  %put DIRECT = &direct PATTERN = &ipat;
%end;

options pageno = 1;
*****;
filename outdoc "H:\dissertn\sas_output\&direct\relate3test.doc";

proc printto new print = outdoc;
run;

proc print data = a_rstats noobs;
  var pattern mean_t01 mean_t02 ac_var cim_vpld cim_vmse t3_var t2_var
    pattern t3trt_05 t2trt_05 qmsp_025 qpasp_025 qmsm_025 qpsm_025;
run;

proc print data = a_relate noobs;
  var PATTERN T3_ST_10 HPLD_10 HMSE_10 CT_GSH_P CT_GSH_M
    GSHGSHPM;
run;

proc print data = a_relate noobs;
  var PATTERN AZ_XT_10 AZ_AT_10 MNQP_10 MNQM_10 CMPLD_SE
    CMMSE_SE RAW_SE;
run;

proc print data = a_relate noobs;
  var PATTERN
    AC_AC_XA AC_GS_XP AC_GS_XM AC_GS_AP AC_GS_AM
    AC_CM_XP AC_CM_XM AC_CM_AP AC_CM_AM
    GS_GS_PM GS_CM_PP GS_CM_PM GS_CM_MP GS_CM_MM
    CM_CM_PM;
run;

```

```

proc print data = a_relate noobs;
  var PATTERN AC_RAW_X AC_RAW_A GS_RAW_P GS_RAW_M
    CM_RAW_P CM_RAW_M;
run;

proc printto;
run;

*****;
proc print data = a_rstats noobs;
  var pattern mean_t01 mean_t02 ac_var cim_vpld cim_vmse t3_var t2_var
    pattern t3trt_05 t2trt_05 qmsp_025 qpsp_025 qmsm_025 qpsm_025;
run;

proc print data = a_relate noobs;
  var PATTERN T3_ST_10 HPLD_10 HMSE_10 CT_GSH_P CT_GSH_M
    GSHGSHPM;
run;

proc print data = a_relate noobs;
  var PATTERN AZ_XT_10 AZ_AT_10 MNQP_10 MNQM_10 CMPLD_SE
    CMMSE_SE RAW_SE;
run;

proc print data = a_relate noobs;
  var PATTERN
    AC_AC_XA AC_GS_XP AC_GS_XM AC_GS_AP AC_GS_AM
    AC_CM_XP AC_CM_XM AC_CM_AP AC_CM_AM
    GS_GS_PM GS_CM_PP GS_CM_PM GS_CM_MP GS_CM_MM
    CM_CM_PM;
run;

proc print data = a_relate noobs;
  var PATTERN AC_RAW_X AC_RAW_A GS_RAW_P GS_RAW_M
    CM_RAW_P CM_RAW_M;
run;

%mend do_all;
*****;

%do_all(S3A);
%do_all(S3H);
%do_all(S3I);
%do_all(S3J);

```

```
%do_all(S3K);  
%do_all(S3L);  
%do_all(S3M);
```

```

*m_ptest2.sas      27nov00;

*This program summarizes the results of the simulations for two centers;
*and the test results from the analyses of the simulations;
*based on the results of preliminary interaction tests;
*Specifically, it summarizes the relationships between positive results;

options mprint mlogic symbolgen notes;
*****;
%macro ptest(pattern);

libname indata "H:\DISSERTN\SAS DATASETS\&direct\P&PATTERN";

data patID;
  length PATTERN $3;
  pattern = "P&PATTERN";
run;

data alltests (keep = nsim T3_ST_10 HPLD_10 HMSE_10 AZ_XT_10 AZ_AT_10
                  MNQP_10 MNQM_10 CMPLDNEG CMMSENEG RAWNEG T3TRT_05
                  T2TRT_05);
  merge indata.allsum indata.allazz indata.acim0102 indata.ags0102;
  by nsim;
run;

data ptest (keep = T3TRT_05 T2TRT_05 P_ST_05 PHMSE_05 PHPLD_05
                  PAZXT_05 PAZAT_05 PMNQP_05 PMNQM_05 PCMPLD05
                  PCMMSE05 P_RAW_05);
  set alltests;

  if T3_ST_10 = 100 then P_ST_05 = T3TRT_05;
  else if T3_ST_10 = 0 then P_ST_05 = T2TRT_05;
  if HPLD_10 = 100 then PHMSE_05 = T3TRT_05;
  else if HPLD_10 = 0 then PHMSE_05 = T2TRT_05;
  if HMSE_10 = 100 then PHPLD_05 = T3TRT_05;
  else if HMSE_10 = 0 then PHPLD_05 = T2TRT_05;
  if AZ_XT_10 = 100 then PAZXT_05 = T3TRT_05;
  else if AZ_XT_10 = 0 then PAZXT_05 = T2TRT_05;
  if AZ_AT_10 = 100 then PAZAT_05 = T3TRT_05;
  else if AZ_AT_10 = 0 then PAZAT_05 = T2TRT_05;
  if MNQP_10 = 100 then PMNQP_05 = T3TRT_05;
  else if MNQP_10 = 0 then PMNQP_05 = T2TRT_05;
  if MNQM_10 = 100 then PMNQM_05 = T3TRT_05;
  else if MNQM_10 = 0 then PMNQM_05 = T2TRT_05;

```

```

        if CMPLDNEG = 1      then PCMPLD05 = T3TRT_05;
    else if CMPLDNEG in (0, 2) then PCMPLD05 = T2TRT_05;
        if CMMSENEG = 1      then PCMMSE05 = T3TRT_05;
    else if CMMSENEG in (0, 2) then PCMMSE05 = T2TRT_05;
        if RAWNEG = 1        then P_RAW_05 = T3TRT_05;
    else if RAWNEG in (0, 2) then P_RAW_05 = T2TRT_05;
run;

proc means data = ptest noprint;
    output out = ptestsum;
run;

data ptestmns (keep = T3TRT_05 T2TRT_05 P_ST_05 PHMSE_05 PHPLD_05
                    PAZXT_05 PAZAT_05 PMNQP_05 PMNQM_05 PCMPLD05
                    PCMMSE05 P_RAW_05);
set ptestsum (keep = _STAT_ T3TRT_05 T2TRT_05 P_ST_05 PHMSE_05
                    PHPLD_05
                    PAZXT_05 PAZAT_05 PMNQP_05 PMNQM_05 PCMPLD05
                    PCMMSE05 P_RAW_05);
where _STAT_ = "MEAN";
run;

data ptestmns;
    merge patID ptestmns;
run;

proc append base = a_ptest data = ptestmns;
run;

%mend ptest;
*****;
*****;
%macro do_all(direct);

proc datasets library = WORK memtype = DATA KILL;
run;

title7 "Comparison of Results Between Methods using Preliminary Testing
      &DIRECT";

%ptest(1);
%ptest(8);
%ptest(2);
%ptest(3);

```



```

%ptest(4);
%ptest(5);
%ptest(11);
%ptest(10);

options pageno = 1;
*****;
filename outdoc "H:\dissertn\sas_output\&direct\ptest2.doc";

proc printto new print = outdoc;
run;

proc print data = a_ptest noobs;
  var PATTERN T3TRT_05 T2TRT_05 P_ST_05 PHMSE_05 PHPLD_05
      PAZXT_05 PAZAT_05 PMNQP_05 PMNQM_05 PCMPLD05
      PCMMSE05 P_RAW_05;
run;

proc printto;
run;
*****;
proc print data = a_ptest noobs;
  var PATTERN T3TRT_05 T2TRT_05 P_ST_05 PHMSE_05 PHPLD_05
      PAZXT_05 PAZAT_05 PMNQP_05 PMNQM_05 PCMPLD05
      PCMMSE05 P_RAW_05;
run;

%mend do_all;
*****;

%do_all(S2A);
%do_all(S2B);
%do_all(S2C);
%do_all(S2F);
%do_all(S2G);

```

```

*m_ptest3.sas      24aug00;

*This program summarizes the results of the simulations of three centers;
*and the test results from the analyses of the simulations;
*based on the results of preliminary interaction tests;
*Specifically, it summarizes the relationships between positive results;

options mprint mlogic symbolgen notes;
*****;
%macro ptest(pattern);

libname indata "H:\DISSERTN\SAS DATASETS\&direct\P&PATTERN";

data patID;
  length PATTERN $3;
  pattern = "P&PATTERN";
run;

data alltests (keep = nsim T3_ST_10 HPLD_10 HMSE_10 AZ_XT_10 AZ_AT_10
                  MNQP_10 MNQM_10 CMPLDNEG CMMSENEG RAWNEG T3TRT_05
                  T2TRT_05);
  merge indata.allsum indata.allazz indata.acim0102 indata.ags0102;
  by nsim;
run;

data ptest (keep = T3TRT_05 T2TRT_05 P_ST_05 PHMSE_05 PHPLD_05
                  PAZXT_05 PAZAT_05 PMNQP_05 PMNQM_05 PCMPLD05
                  PCMMSE05 P_RAW_05);
  set alltests;

  if T3_ST_10 = 100 then P_ST_05 = T3TRT_05;
  else if T3_ST_10 = 0 then P_ST_05 = T2TRT_05;
  if HPLD_10 = 100 then PHMSE_05 = T3TRT_05;
  else if HPLD_10 = 0 then PHMSE_05 = T2TRT_05;
  if HMSE_10 = 100 then PHPLD_05 = T3TRT_05;
  else if HMSE_10 = 0 then PHPLD_05 = T2TRT_05;
  if AZ_XT_10 = 100 then PAZXT_05 = T3TRT_05;
  else if AZ_XT_10 = 0 then PAZXT_05 = T2TRT_05;
  if AZ_AT_10 = 100 then PAZAT_05 = T3TRT_05;
  else if AZ_AT_10 = 0 then PAZAT_05 = T2TRT_05;
  if MNQP_10 = 100 then PMNQP_05 = T3TRT_05;
  else if MNQP_10 = 0 then PMNQP_05 = T2TRT_05;
  if MNQM_10 = 100 then PMNQM_05 = T3TRT_05;
  else if MNQM_10 = 0 then PMNQM_05 = T2TRT_05;

```

```

        if CMPLDNEG in (1, 2) then PCMPLD05 = T3TRT_05;
    else if CMPLDNEG in (0, 3) then PCMPLD05 = T2TRT_05;
        if CMMSENEG in (1, 2) then PCMMSE05 = T3TRT_05;
    else if CMMSENEG in (0, 3) then PCMMSE05 = T2TRT_05;
        if RAWNEG in (1, 2) then P_RAW_05 = T3TRT_05;
    else if RAWNEG in (0, 3) then P_RAW_05 = T2TRT_05;
run;

proc means data = ptest noprint;
    output out = ptestsum;
run;

data ptestmns (keep = T3TRT_05 T2TRT_05 P_ST_05 PHMSE_05 PHPLD_05
    PAZXT_05 PAZAT_05 PMNQP_05 PMNQM_05 PCMPLD05
    PCMMSE05 P_RAW_05);
set ptestsum (keep = _STAT_ T3TRT_05 T2TRT_05 P_ST_05 PHMSE_05
    PHPLD_05
    PAZXT_05 PAZAT_05 PMNQP_05 PMNQM_05 PCMPLD05
    PCMMSE05 P_RAW_05);
where _STAT_ = "MEAN";
run;

data ptestmns;
    merge patID ptestmns;
run;

proc append base = a_ptest data = ptestmns;
run;

%mend ptest;
*****;
*****;
%macro do_all(direct);

proc datasets library = WORK memtype = DATA KILL;
run;

title7 "Comparison of Results Between Methods using Preliminary Testing
    &DIRECT";

%do ipat = 1 %to 11;
    %ptest(&ipat);
%end;

```

```

options pageno = 1;
*****;
filename outdoc "H:\dissertn\sas_output\&direct\ptest3.doc";

proc printto new print = outdoc;
run;

proc print data = a_ptest noobs;
  var PATTERN T3TRT_05 T2TRT_05 P_ST_05 PHMSE_05 PHPLD_05
      PAZXT_05 PAZAT_05 PMNQP_05 PMNQM_05 PCMPLD05
      PCMMSE05 P_RAW_05;
run;

proc printto;
run;
*****;
proc print data = a_ptest noobs;
  var PATTERN T3TRT_05 T2TRT_05 P_ST_05 PHMSE_05 PHPLD_05
      PAZXT_05 PAZAT_05 PMNQP_05 PMNQM_05 PCMPLD05
      PCMMSE05 P_RAW_05;
run;

%mend do_all;
*****;

%do_all(S3A);
%do_all(S3H);
%do_all(S3I);
%do_all(S3J);
%do_all(S3K);
%do_all(S3L);
%do_all(S3M);

```

BIBLIOGRAPHY

- Azzalini, A. & Cox, D. R. (1984). Two New Tests Associated with Analysis of Variance. J.R.Statist Soc.B, 46, 335-343.
- Bancroft, T. A. (1964). Analysis and Inference for Incompletely Specified Models Involving the Use of Preliminary Test(s) of Significance. Biometrics 427-442.
- Center for Drug Evaluation and Research, Food and Drug Administration, U.S. Department of Health and Human Services. Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications. July 1988.
- Ciminera, J. L., Heyse, J. F., Nguyen, H. H., & Tukey, J. W. (1993). Tests for Qualitative Treatment-By-Centre Interaction Using A 'Pushback' Procedure. Statistics in Medicine, 12, 1033-1045.
- Fleiss, J. L. (1986). Analysis of Data from Multiclinic Trials. Controlled Clinical Trials, 7, 267-275.
- Gail, M. & Simon, R. (1985). Testing for Qualitative Interactions between Treatment Effects and Patient Subsets. Biometrics, 41, 361-372.
- Gallo, P. P. (1998). Practical Issues in Linear Models Analyses in Multicenter Clinical Trials. Biopharmaceutical Report, 6, 1-9.
- Goldberg, J. D. & Koury, K. J. (1990). Design and Analysis of Multicenter Trials. In D.A.Berry (Ed.), Statistical Methodology in the Pharmaceutical Sciences (Marcel Dekker.
- Gould, A. L. (1998). Multi-Centre Trial Analysis Revisited. Statistics in Medicine, 17, 1779-1797.
- Hochberg, Y. & Tamhane, A. C. (1987). Multiple Comparison Procedures. New York: John Wiley and Sons, Inc.
- Hsu, J. C. (1996). Multiple Comparisons: Theory and Methods. London: Chapman and Hall.
- International Conference on Harmonisation (1996). Structure and Content of Clinical Study Reports. Federal Register, 61, 37320-37343.

- International Conference on Harmonisation (1998). Guidance on Statistical Principles for Clinical Trials. Federal Register, 63, 49583-49598.
- Jones, B., Teather, D., Wang, J., & Lewis, J. A. (1998). A Comparison of Various Estimators of a Treatment Difference for a Multi-Centre Clinical Trial. Statistics in Medicine, 17, 1767-1777.
- Källén, A. (1997). Treatment-by-Center Interaction: What is the Issue? Drug Information Journal, 31, 927-936.
- Lewis, J. A. (1995). Statistical Issues in the Regulation of Medicines. Statistics in Medicine, 14, 127-136.
- Lin, Z. (1999). An Issue of Statistical Analysis in Controlled Multicenter Studies: How Shall We Weight the Centers? Statistics in Medicine, 18, 365-373.
- Nelder, J. A. (1994). The Statistics Of Linear Models: Back to the Basics. Statistics and Computing, 4, 221-234.
- Nelder, J. A. & Lane, P. W. (1995). The Computer Analysis of Factorial Experiments: In Memoriam -- Frank Yates. The American Statistician, 49, 382-385.
- Overall, J. E. (1979). General Linear Model Analysis of Variance. In J. Levine (Ed.), Coordinating Clinical Trials in Psychopharmacology: Planning, Documentation, and Analysis. Washington, D.C.: U.S. Government Printing Office.
- SAS Institute Inc (1997). SAS/STAT Software: Changes and Enhancements through Release 6.12. Cary, NC, USA: SAS institute Inc.
- Senn, S. (1998). Some Controversies in Planning and Analysing Multi-Centre Trials. Statistics in Medicine, 17, 1765.
- Snapinn, S. M. (1998). Interpreting Interaction: The Classical Approach. Drug Information Journal, 32, 433-438.
- Snedecor, G. W. & Cochran, W. G. (1967). Statistical Methods. (6 ed.) Ames, Iowa, U.S.A.: The Iowa State University Press.
- Winer, B. J. (1971). Statistical Principles in Experimental Design. (2 ed.) New York: McGraw-Hill.
- Yates, F. (1934). The Analysis of Multiple Classifications with Unequal Numbers in the Different Classes. Journal of the American Statistical Association, 29, 51-66.