Title

Estimating treatment effect in the presence of noncompliance measured with error

Author

Leslie, Anne Kenna

Publication Date

2001

Estimating Treatment Effect in the Presence of Noncompliance Measured with Error: Power, Precision, and Robustness of Data Analysis Methods

by

Leslie Anne Kenna

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Chemistry

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

For my mother, Dorothy Kenna—

a person who relentlessly reminds

"take all of your medicine, even if you feel better",

but compassionately serves

orange-flavored aspririn and cherry-tasting cough syrup.

ACKNOWLEDGEMENTS

I've been told that it's time to graduate when your thesis advisor can't teach you anything else. This is supposed to happen in four to five years. Based on this criterion, having Lewis Sheiner as an advisor makes it difficult to ever feel ready (or willing) to leave! Much as I hoped that I would wake up one morning and have his sharp insight, some things just can't be taught. I consider it an accomplishment to now be able to follow him in discussion!

Intellectually and spiritually, learning to think statistically is as significant as learning to read. At least this is how it feels having been trained by Lewis. Perhaps more important than creating scientists, Lewis demonstrates how to be a mentor. The learning process delights Lewis. At a moment's notice, he sets aside whatever he is working on to discuss ideas for as long as his students and fellows ask. Nothing sits on his desk unread for more than a day, and he really thinks about what you write! He encourages his students and fellows to collaborate with others and proudly introduces us to his peers. He gives everyone credit for their work. The only unethical thing he does is give you credit for his ideas, too. Thankfully, his support, advice, and statistical consultation will always be just an email message away.

I am deeply grateful to both Lewis and his wife, Michael, for their understanding of and insight on the personal challenges of graduate school.

I owe tremendous thanks to Les Benet. As with everything, he made it possible for me to work with Lewis with one enthusiastic phone call. Les has been a valuable guide during every stage of my graduate career. When I first started out, his extraordinary optimism and sense of humor made me feel that I belonged at UCSF. During my oral qualifying exam, thesis committee meetings, and seminars, Les' questions and feedback challenged me to defend and rethink my work. His detailed comments significantly improved my thesis. Although Les's lab is large enough to be considered a biotech company, he genuinely cares about the future of all students in the department, regardless of whether they are in his lab or not. The only unethical thing Les does is write excessively kind letters of recommendation.

I am grateful to the other members of my thesis committee. Davide Verotta is a friendly and approachable mentor whom I admire for both his scientific and musical talents. Steve Shiboski was very generous with his time, encouragement, and input on my project.

I want to thank Kathy Giacomini for her advice on choosing a thesis advisor and helping me prepare for my oral qualifying exam. I want to thank Terry Blachke for his help during the preparation for my oral qualifying exam, as well.

I want to thank my academic advisor, Tony Hunt, for his advice on how to efficiently jump through miscellaneous bureaucratic hoops, for giving me detailed feedback on my oral presentations, and for many interesting discussions regarding future directions for my project. He is a trusted advocate for students and greets all ideas, regardless of how crazy they may sound, not only with encouragement but also with a list of contacts that can help make them happen.

I owe thanks to the U. S. Pharmacopoeia for supporting my research, in part.

I want to thank the mentors who have given me so much of their time, despite having no formal responsibility for my education. I consider myself fortunate that I know

Harriet Hopf, a UCSF Professor of Anesthesiology, personally, and that I didn't just read about her when she was named "Advisor of the Year" the first year the award was given! It amazes me how she manages to balance time in the operating room, a research lab, family life, and a passion for world travel, while keeping tabs on the personal and professional lives of all of the students and residents she mentors.

I want to thank Laura Edsberg for giving me as a job as a lab assistant in college and for continuing to stay in touch with me from 2000 miles away. She was the first person to show me that working in a lab can be fun. She, more than anyone else, prepared me for what graduate school would entail.

I am grateful to have met two people at UCSF who I consider both friends and mentors. My work progressed very much during Niclas Jonsson's postdoctoral fellowship in the lab. He brought exactly what I needed—Splus expertise and a sense of humor. Rae Yuan, my assigned departmental peer mentor, helped me navigate through graduate school when she was at UCSF and now gives me trusted career advice.

I want to thank Stuart Beal for many interesting discussions, especially the ones about public radio, the *Synapse*, and Halloween costumes. His sense of humor made time spent sitting outside of his office a real treat.

One fringe benefit of working with Lewis is that he attracts a diverse and wonderful group of people to the lab. During the time I've spent here, French, Swedish, Dutch, Italian, German, Spanish, Japanese, Chinese, and Korean have been spoken in the lab. I appreciate the feedback I've gotten on my work from labmates and I thank them for attending my PC 220 seminars. I want to thank Marta Valle and Marc Pfister for getting the group to do things socially. I want to thank Jianfeng Lu for making us laugh. I want to thank Line Labbe for being such a wonderful person to talk to. I want to thank Liping Zhang for being like a sibling in the lab. I want to thank Dolors for showing me how much fun it is to collaborate on a project and thank Micha Levi for solving many frustrating computer problems.

I don't know where to acknowledge Barbara Weichmann. We met because she is the lab's administrative assistant. But our relationship has developed in other unexpected ways. She is an inspirational yoga instructor. She is a trusted friend. We became family after I posed as her son's long-lost cousin at his wedding.

I have made many wonderful friends during my time at UCSF. Shawn Flanagan was not only a great person to be partnered up with to teach Pharmacokinetics 128 & 129, but he and his wife, Debra, are valued friends. Carolina Reyes, thankfully, taught me to live and cook like a Californian. Karin Gerstin and Mark Dresser always make me laugh. My roommates, Linda and Sharon, made it nice to come home every day.

I owe my sanity to my old friends Timmy Pettolina, Jennifer Murphy, and Ignacio Manso.

I want to thank the Graduate Students' Association and the staff at the *Synapse* for giving me a place to unwind on campus. I want to thank the staff of the Science and Health Education Partnership (SEP) for their commitment to science education at all grade levels. In training me to run an after-school science club for middle school girls, the SEP staff showed me how to get over my own lack of confidence as a scientist. I want to thank the people of KUNC-FM in Greeley, Colorado for giving me the best summer job I have ever had in my life! I want to thank my creative writing group for sharing their work, reading my work, and keeping a sense of humor about the whole thing.

I couldn't have gotten into graduate school or finished my thesis without the unconditional love, understanding, and support of my family. It's impossible to list what they've given me! Perhaps the most important lesson my parents have taught me is that success can't be measured with letters or numbers. The most important gift they've given me is a sense that I can call them any time of day with any request.

My two sisters, Susan and Elaine, helped me keep my sense of humor through this entire process. They are not only great friends, but also role models. Both are successful career women who also manage to have a rich family life. Because they are great people, they married great people. Chris Boyle, Captain of Engine Company 94 in the Bronx, was my hero long before September 11th. My thesis work pales in comparison to the work Joe Turner did restoring his 150+ year old house, but it's the closest analogy to research I've found! The process is slow and full of surprises—much like when he ripped up the floorboards and realized that a tree trunk shaped like a bear was supporting the entire foundation!

I want to thank my nieces and nephews—Aidan Turner, Ariel Boyle, Ryan Turner, Sarah Boyle, and Timmy Turner—for their love and enthusiasm. They make me feel cool for being a scientist; even though I don't do anything that they consider scientific work ("a scientist studies leafs, trees, flowers, dinosaurs, dinosaur bones, rocks, fossils, and magnifying glasses.")

Last, but not least, I want to thank my fiancé, Noel Southall, for being such a sweet, funny, and patient guy and for wanting to share everything with me. It snowed in San Francisco after our first date—one of five times during the past century. That was only the beginning of the beauty and surprises with him. We rarely talk about science, but

I consider it a bonus that a person who makes life so much fun can also understand my work!

Leslie Kenna

December 2001

ABSTRACT

Estimating Treatment Effect in the Presence of Noncompliance Measured with Error: Power, Precision, and Robustness of Data Analysis Methods

Leslie Anne Kenna

Fractional compliance (C) with the assigned, or nominal, dose (D_n) gives rise to unintended variability in exposure (D) during confirmatory clinical trials. If C is not a confounder, that is, C only influences response through its influence on drug exposure, then in principle, assuming a model $D=f(C,D_n)$ allows one to estimate the exposure-response relationship $P(Y|D)$.

The measurement of C presents many challenges. The most accurate measures of compliance are often the least feasible to obtain. Given that biased compliance questionnaire assessments (C_Q; $C_Q \geq C$) are available in all N subjects enrolled in a clinical trial, but accurate compliance measures (C_M; $C_M=C$), for example, from electronic medication caps which record openings, are only known in a random fraction of N, how does one estimate $P(Y|D)$? Simulation studies are performed to compare several analysis methods in terms of their precision for estimates of $P(Y|D)$ and their power to reject $P(Y|D>0) = P(Y|D=0)$. A "maximum likelihood" (ML) method, which uses all C_Q,C_M,Y data, and calibrates C_Q to C_M is compared to other methods, which use only one, or both, or neither of C_M and C_Q but do not calibrate (neither = Intention-to-Treat (ITT) which assumes C=100% in all).

Given that the key assumptions of **ML** are met (***A3***: C_M accurately measures C, and ***A2***: M is assigned at random), **ML** yields the most precise estimates of P(Y|D) over widely varying clinical trial designs, extremes in quality and quantity of compliance information, and a range of drug effect sizes. **ML** is most beneficial given data sets having sparse compliance information. However, **ITT** can be just as powerful as **ML**. **ML** maintains its superior precision when ***A3*** is violated and is equivalent to the best performing methods when ***A2*** is violated. When ***A3*** and ***A2*** are violated simultaneously, **ML** has the second best performance. The relative performance of all methods is maintained when a real data set is analyzed.

In conclusion, **ML** is an efficient and robust method for determining P(Y|D) given a trial with compliance measured via a calibration design.

12/20/01

TABLE OF CONTENTS

LIST OF TABLES

Chapter 5

Chapter 6

Chapter 8

Appendix

LIST OF FIGURES

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Chapter 9

Appendix

ABBREVIATIONS AND SYMBOLS

Complete Data for a Generic Individual in a Clinical Trial

Y = Pharmacodynamic response

C = True compliance

C_M = Compliance measured with an electronic monitor (“$_M$” = Monitor)

C_Q = Compliance measured via patient self-report (“$_Q$” = Questionnaire)

D_n = Assigned dose (“$_n$” = nominal)

D = $D(C,D_n)$ = Exposure

M = Indicator variable: electronically monitored (M=1) / not monitored (M=0)

Note:

1. Any of above, when subscripted using the index i, explicitly denotes the i^{th} individual's value.
2. For random variables a and b,
    a. P(a): the density or probability mass function of a
    b. P(a|b): the conditional density of a given b
    c. $a \perp b$: a is distributed independently of b.
    d. “^” denotes an estimate

Experimental Set Up

N	Number of subjects in the clinical trial
N_M	Number of subjects with M=1; $N_M \leq N$

f_M	Fraction of subjects with M=1; $f_M = N_M/N$
Y, M, D_n	Binary
C, C_M, C_Q, D	Categorical (3 levels)
Y, C_Q, D_n and M	Observed in all
C_M	Observed in subjects with M=1
C	Unobserved

Simulation Models

N_{lhs}	The number of sets of parameter values drawn by Latin Hypercube Sampling
$D(C,D_n)$	Function mapping compliance and nominal dose (C, D_n) to exposure (D); for more details refer to "Simulation Design For Each Clinical Trial" in section A.1 of the Appendix

Simulation Microparameters

ρ	Logit{P(Y|D=0)}
a, *b*	Parameters of $P(C_Q\|C)$

Model Parameters

Estimated

θ	All parameters to be estimated; $\theta = \theta_1, \theta_2$ for the ML method of analysis, $\theta = \theta_1$ for all other methods of analysis
θ_1	Parameters of the model P(Y|D)

θ_2	Parameters of the model $P(C\|C_Q)$

Fixed Constants

θ_{1prior}	Prior mean of θ_1
θ_{2prior}	Prior mean of θ_2
λ_1	Weight of the prior penalty for $\theta_1 \neq \theta_{1prior}$
λ_2	Weight of the prior penalty for $\theta_2 \neq \theta_{2prior}$

Methods of analysis

ALL	C_M is known in "All" subjects (N subjects) and used to estimate θ
BA	Uses "Best Available" compliance data to estimate θ
CD	"Complete Data"; uses only C_M data in M=1 to estimate θ
ITT	"Intention-To-Treat"; Assumes C=100% to estimate θ
ML	"Maximum Likelihood" Estimator; uses all C_M and C_Q data to estimate θ
BSR	"Believe Self-Report"; uses only C_Q data to estimate θ

SECTION I

GENERAL INTRODUCTION

Chapter 1: Patient Compliance During Clinical Trials and its Influence on the Estimated Drug Effect

1.1 Overview of Compliance

1.1.1 Definition

Neglecting to take medication as prescribed is a major cause of variability in drug exposure and has been associated with the failure of many treatments (Didlake, Dreyfus et al. 1988; Bond and Hussar 1991; Cramer and Spilker 1991; Harter and Peck 1991; Urquhart 1992; Urquhart 1997; Kastrissios and Blaschke 1998). Compliance, a term used to describe the agreement between a patient's actual drug taking and the prescribed regimen(Urquhart 1992; Urquhart 1994), is not a new problem in therapeutics. Records documenting physician concern about patient compliance date to the time of Hippocrates (Didlake, Dreyfus et al. 1988; Bond and Hussar 1991; Cramer 1991; Ickovic and Meisler 1997).

1.1.2 Measurement

Centuries ago, physicians may have asked patients about their compliance or inferred it from their response to treatment. Today, investigators still elicit patient self-reports of compliance(de Klerk, van der Heijde et al. 1999; Chesney, Ickovics et al. 2000), but may perform pill counts(Lee, Kusek et al. 1996), check plasma levels(Maenpaa, Javela et al. 1987; Hardy, Kumar et al. 1990; Kapur, Ganguli et al. 1991; Bloch, Gur et al. 1992), or monitor medication bottle opening and closing with an electronic device wired into the cap(Cramer, Mattson et al. 1989; Urquhart 1997), as well. Each offers a different approach to quantifying drug intake.

A pill count compares the number of pills in the patient's possession at the end of a dosing period to the amount that should have been remaining if compliance were perfect. Self-report tools, such as diaries and questionnaires, ask the patient how many pills they recall taking (or missing) during an interval of time. Biological assays infer compliance through the presence of drug, metabolite, or other intake markers in bodily fluids. Electronic chips in a medication bottle cap continuously record the time of pill bottle opening and closing. Presumably, an opening and closing event signals that the appropriate dose has been ingested. Biological assays and a clinician's observation of the patient ingesting medication, known as directly observed therapy (Weis, Slocum et al. 1994), are considered direct measures of compliance. Self-report, pill counts, and electronic monitors are labeled indirect measures (Farmer 1999).

1.1.3 Metrics

Numerous approaches to transforming the data collected using compliance measuring tools into a univariate summary of intake have been reported (Vrijens and Goetghebeur 1997). Percent compliance is most commonly used and most often defined as the fraction of prescribed doses taken during some interval of intake observation. That interval may be the entire duration of dosing, or, perhaps, just a few days prior to a visit with the clinician. The duration of time is dictated by the technique used to measure compliance.

For example, an observed drug level reflects compliance over the compound's previous four half lives (Urquhart 1997) while pill counts reflect compliance over the entire period of time between counts. Self-reported compliance can reflect intake over the entire study duration if the patient uses a diary to record compliance daily. Due to the limitations imposed by memory, when questionnaires are used, subjects are usually asked to recall their intake just a few days prior to visiting the clinician. Electronic caps monitor compliance continuously, so there is no technically imposed limit on the duration of time over which percent compliance can be measured with this tool.

In addition to percent compliance, other compliance metrics reported in the literature include "percentage of prescribed dosing days with the correct intake"(Vrijens and Goetghebeur 1997), (which is equivalent to "compliance rate"(Cramer, Mattson et al. 1989)) "therapeutic coverage"(Detry 1994; Meredith and Elliott 1994; Urquhart 1994; Meredith 1999), frequency of "drug holidays"(Urquhart and Chevalley 1988), and various metrics for describing dose timing—variability around the median dosing time, percentage of too short or too long dosing intervals, and median and quantiles of the dosing intervals (Vrijens and Goetghebeur 1997).

Percent dosing days refers to the percent of days on which the patient took the prescribed number of pills. Therapeutic coverage refers to the fraction of time during which a patient ingests sufficient medication to keep drug concentrations above some minimum efficacious level. This is related to a drug's "forgiveness", or the period of time during which an effect persists despite the absence of drug in the measurement compartment (Urquhart 1997). A drug holiday was originally defined as a discontinuation of drug intake for three or more consecutive days (Urquhart and Chevalley 1988;

Urquhart and De Klerk 1998). Others later defined it as one or more days without drug intake (Vrijens and Goetghebeur 1997).

Self-reported compliance is often quantified using a coarse description of dosing (discrete categories) rather than using continuous values on a 0-100% scale. Labeling patients as compliant, non-compliant, or moderately compliant has intuitive appeal for the clinician. In contrast, some have culled a multivariate description of drug intake from compliance records. The entire time series of bottle opening and closing events has been captured with several parameters by modeling the data as a Markov process(Girard, Blaschke et al. 1998).

One goal of developing such summary statistics is to identify predictors of patient dosing behavior. For example, the Markov model reveals that the day of the week correlates with percent compliance—patients tend to skip more pills and take morning doses later on the weekend compared to weekdays(Girard, Blaschke et al. 1998). Predictors of compliance tend to be factors such as dosing frequency, dietary restrictions, pill burden, and patient-provider relationships(Chesney 2000), while sociodemographic variables are not predictive of one's drug intake behavior(Chesney 2000; Wright 2000). For example, compliance, defined as the percent of days on which the medication bottle was opened, was not influenced by age, sex or nationality, in a clinical trial comparing the cardiovascular effects of aspirin to placebo(Waeber, Leonetti et al. 1999). The Markov model, however, identified age as a predictor of compliance(Girard, Blaschke et al. 1998). Perhaps this reflects the model's ability to capture dose amount and timing information.

Wide intra-(Cramer, Scheyer et al. 1990; Waeber, Leonetti et al. 1999) and inter-individual(Kastrissios and Blaschke 1997) variability in compliance has been observed in controlled clinical trials. Urquhart compared reported values of electronically monitored compliance for three different drugs—therapies for glaucoma, epilepsy, and arthritis¯and noted that the distribution of patient compliance appears similar across diverse medical conditions(Urquhart and De Klerk 1998). Patients took an average of 76% (range: 0-100%)(Kass, Meltzer et al. 1986) of prescribed topical pilocarpine for glaucoma, 76% (range: 30-100%)(Cramer, Mattson et al. 1989) of an oral epileptic, and 81% of a non-steroidal anti-inflammatory drug (range: 10-100%)(de Klerk and van der Linden 1996). The similarity between compliance distributions for a wide variety of ambulatory patients suggests that patient compliance is more related to behavioral qualities, rather than pathophysiological conditions. It suggests why an individual's compliance is difficult to predict(Kastrissios and Blaschke 1997).

There are exceptions to this rule. Compliance may be lower with drugs to which patients can attribute unpleasant side effects(Chesney 2000). An example is cholestyramine—a drug whose side effects include gastrointestinal discomfort. Patients who received the drug during clinical trials were less compliant than patients who received placebo(Program 1984; Urquhart 1991).

Note that a doctor's intuition about compliance is poorly predictive of compliance(Kass, Gordon et al. 1986; Turner and Hecht 2001). Physicians have been shown to correctly label patients as compliant or noncompliant only 1 out of every 2 times—as poor as flipping a coin to decide(De Geest, Borgermans et al. 1995; Rich, Gray et al. 1996). One study revealed that provider estimates of compliance explain only 26%

(95% CI, 6%-47%) of the variation in pill count adherence, while patient self-report explains 72% (95% CI, 52%- 96%)(Bangsberg, Hecht et al. 2001).

1.2 Compliance and Drug Effect

1.2.1 Influence of Compliance on Pharmacodynamic Response to Efficacious Drugs

It is difficult to say whether failing to comply would have helped or harmed one's health in the early days of medicine. However, safe and effective medications cannot work in people who do not take them(Koop 1984). There is considerable evidence that forgetting to take several doses of an immunosuppressive results in the rejection of a transplanted organ. One group reported that noncompliance accounts for 13% of graft loss(Hong, Sumrani et al. 1992) and increases to 27.6% of graft loss 2-3 years post-transplantation(Dunn, Golden et al. 1990). The level of compliance among transplant recipients is less than ideal. In one study, 22% of 148 adult renal transplant recipients admitted missing several doses each month during the past year(De Geest, Borgermans et al. 1995).

For some medications, there are risks associated with skipping pills beyond the anticipated loss of drug effect. It may be equally dangerous to self-medicate intermittently than to take no drug at all! Patients are prone to develop a hypertensive crisis after several doses of a β–blocker are missed(Urquhart 1997). (Upregulation of beta-adrenergic receptor production in patients exposed to β–blockers is the proposed mechanistic basis for this effect. The degradation of β-receptors occurs over a longer period of time than the duration over which patients may alter their β–antagonist plasma levels through self-medication. When patients skip several doses in a row, the resultant drop in drug level increases the fraction of unbound β–receptors. This sudden change in free β–receptor level enhances sensitivity to endogenous β–agonists.) Psaty and coworkers report a fourfold increase in the relative risk of coronary heart disease in subjects whose record of beta blocker prescription filling revealed less than 80% compliance(Psaty, Koepsell et al. 1990).

More insidious are the public health risks of noncompliance, as evidenced by the well-known example of incomplete dosing with antibiotics. Drug-resistant infectious strains are likely to proliferate when patients fail to self-medicate above the minimum effective concentration for an adequate period of time(Lipsitch and Levin 1998; Mitchison 1998). Lack of efficacy and the emergence of drug-resistant strains of HIV have been linked to non-adherence with antiretroviral therapy(Chesney, Ickovics et al. 1999). Compliance is a particularly thorny issue for patients taking anti-HIV medications due to the complexity of the regimens prescribed(Chesney, Morin et al. 2000). Only half of patients take all antiretroviral medication in accordance with time and dietary instructions in a given week(Nieuwkerk, Sprangers et al. 2001).

1.2.2 Compliance and Confirmatory Clinical Trials

Compliance in clinical trials is as much a determinant of outcome as in clinical practice(Peck 1999). The standard clinical trial analysis method, the intention-to-treat (**ITT**) procedure, estimates drug efficacy by pooling the outcomes of patients who are assigned to drug but do not take it compared with the outcomes of those who are fully compliant. Because of this, **ITT** is said to estimate 'use effectiveness'(Sheiner and Rubin 1995). If all patients are perfectly compliant with the prescribed treatment, **ITT** estimates the true pharmacologic effect of drug, or 'method effectiveness'. If patients are less than perfectly compliant with an effective therapy, the **ITT** approach yields a downwardly biased estimate of method effectiveness and can possibly impact the result of a clinical trial(Kastrissios and Blaschke 1997; Hasford 1999). Statistically speaking, poor compliance increases **ITT**'s chance of failing to reject the null hypothesis when it should be rejected. Some consider drugs to be mislabeled if **ITT** average values for drug efficacy are offered as the only dosing guidelines(Lasagna and Hutt 1991).

The Food and Drug Administration (FDA) requires an **ITT** analysis for the determination of efficacy from confirmatory clinical trials since it lends itself to a causal interpretation of the outcome(Peck 1999). The **ITT** estimate is causal as it estimates the difference in average response caused by the difference in randomly assigned dose (not ingested dose).

1.3 Approaches That Reduce the Influence of Compliance on Estimated Drug Effect

1.3.1 Alter Clinical Trial Design

Under the **ITT** paradigm, investigators may try to protect study power by increasing the number of subjects enrolled(Freedman 1990). They may use poor compliance as an exclusion criterion or intervene to improve compliance so **ITT**'s estimate of use effectiveness approaches true method effectiveness. Each of these solutions has caveats. Obviously, all three approaches tax available resources.

Increasing study size and excluding poor compliers are only options if there is a large enough patient pool to draw from. Since there are no reliable predictors of compliance(Lerner, Gulick et al. 1998; Wright 2000), noncompliers must be identified for exclusion by performing a run-in, or mock, trial before the study commences. The run-in may use a placebo, however, compliance with placebo does not necessarily predict an individual's compliance with an active compound(Sheiner and Rubin 1995). If a run-in is performed using the drug, one runs the risk that subjects may become unblinded or exhibit crossover effects. The run-in, therefore, threatens the generality of the clinical trial(Pablos-Mendez, Barr et al. 1998). Furthermore, it is unclear what constitutes an adequate run-in duration to identify the poor compliers(Kastrissios and Blaschke 1997). Misclassification of noncompliers as compliers during the run-in decreases the efficiency of the clinical trial(Brittain and Wittes 1990).

Interventions to improve compliance involve alerting patients to take drug at each dosing event or counseling behavioral modifications that enable patients to self-medicate(Haynes 2000). Systems for alerting patients range from notification via email and pagers to having support staff telephone the subject(Urquhart 1997). The most extreme form of intervention is directly observed therapy (DOT)(Barker and Millard 2000; Volmink and Garner 2001). DOT requires subjects to visit the study site to receive treatment at every dosing event.

An analysis of 19 randomized controlled trials designed to measure the efficacy of interventions to improve compliance revealed that 17 were successful. They involved some combination of more convenient care, dissemination of drug information, patient counseling, periodic reminders, self-monitoring, clinician reinforcement, family therapy and additional supervision or attention(Haynes 2000). Some data suggest that these approaches are only successful in improving adherence while the intervention is applied(Cummings, Becker et al. 1981). Patients receiving directly observed anti-HIV therapy on weekdays and self-administering doses on weekends had significantly reduced intake on Saturdays and Sundays(Wall, Sorensen et al. 1995). Selecting for a group of perfect compliers is desirable because it allows one to estimate the physiologic effect of the dose tested. However, since compliance is variable under conditions of real use, a trial of flawless compliers may reveal little about how the drug will perform in the clinic.

1.3.2 Alter Data Analysis

An appealing approach—because it requires a minimal amount of resources and yields information most relevant in practice—is to perform an as-treated analysis of confirmatory trial data to supplement the ITT approach. More specifically, the proposal is to measure compliance to determine actual drug exposure, which is then related to pharmacological response. This has been referred to as treating compliance as causing a natural experiment in dose ranging(Urquhart and Chevalley 1988).

1.3.2.1 Caveat #1: Confounding

Despite the potential benefits, this approach has rarely been used in developing dosing guidelines(Peck 1999). An as-treated analysis using compliance data poses two serious data analytic challenges. The first is an issue of confounding and the second is an issue of compliance measurement.

The issue of confounding arises because a subject's compliance is unknown at the outset of a trial, and, therefore, is, technically, an outcome of the treatment. (It cannot be a variable on which stratification occurs since, as noted earlier, the search for individual predictors of compliance has been fruitless to date(Kastrissios and Blaschke 1997)).

Using compliance information to determine exposure for exposure-response estimation effectively treats compliance as an independent variable in data analysis. The extent to which a subject's intake causes his pharmacodynamic response through drug exposure versus the possibility that both response and compliance are driven by another factor is unknown. Without additional data or assumptions, the estimated exposure-response relationship may be biased. Several model-based approaches to determining exposure-response when treatment taken differs from treatment assigned have been reported(Efron and Feldman 1991; Sheiner and Rubin 1995; Angrist, Imbens et al. 1996; Goetghebeur, Molenberghs et al. 1998; Robins 1998).

The approaches to finding causal estimates of exposure when there is confounding have rested on an assumption about the relationship between compliance in drug and placebo groups(Efron and Feldman 1991; Angrist, Imbens et al. 1996; Goetghebeur, Molenberghs et al. 1998), rested on an assumption about response in non-compliers(Sheiner and Rubin 1995), or taken advantage of hypothetical patient covariates to stratify on compliance(Robins 1998).

The relationship between compliance to drug and compliance to placebo must be specified as the proper control for the subjects who comply with drug is not necessarily the group of subjects who comply with placebo. It is the subset assigned to placebo that would have been compliant with drug had they been assigned to that study arm. Note that having the same distribution of compliance in the drug and placebo groups, theoretically, does not reduce the need for such an assumption. (For example, patients in a clinical trial comparing the cardiovascular effects of aspirin had the same mean and variance in compliance regardless of being assigned to drug or placebo(Waeber, Leonetti et al. 1999)). Only measuring compliance with drug and placebo in the same individual does.

Efron and Feldman (1991) developed a causal estimator of the effect of exposure to the lipid-lowering drug, cholestyramine, on coronary heart disease using data collected during the Lipid Research Clinics Coronary Primary Prevention Trial, or LRCPPT (Program 1984). The LRCPPT data set received much attention because it suggests there is confounding between compliance and response; a trend between compliance and response was observed in both the treatment and placebo groups. Furthermore, subjects were observed to have lower compliance with drug than placebo, complicating the task of finding the proper control for subjects assigned to dose in the placebo group.

Efron and Feldman (1991) note that the steep compliance-response relationship observed in subjects assigned to drug and the shallow compliance-response relationship observed in subjects assigned to placebo is evidence of a dose-response relationship. Their strategy is to recover the dose-response relationship from the compliance-response relationship by estimating the difference in response for those assigned to drug and those assigned to placebo at matched levels of compliance. The authors assume: (1) there is no difference in response between 0% compliers to drug and 0% compliers to treatment, and (2) compliance is an inherent attribute of the patient ("perfect blind assumption"), which allows them to write a model relating an individual's compliance with drug to his compliance with placebo. Their results are likely sensitive to these assumptions—it has been demonstrated that incorrectly assuming that compliers to placebo are the proper control group for compliers with drug leads to severe bias in estimates of drug effect(Albert and Demets 1994).

Sheiner and Rubin's (1995) approach is an example of a generic modeling method known as instrumental variables. Their analysis rests on two assumptions—first, the decision to comply or not comply occurs early in the trial, which provides a basis for believing the second, and key, assumption that outcomes in drug noncompliers are the same as they would have been had the non-compliers been assigned to the control treatment. Under this scenario, only the marginal distributions of compliance to placebo and compliance to drug are required to yield an unbiased estimate of the causal relationship between exposure and response. However, as the authors point out, this approach requires more investigation to extend to applications beyond the analysis of vaccine trial designs. Note that vaccine trial designs were used as an example for which the key assumptions are valid: (1) no drug is available to those who are not assigned to receive it, (2) subjects have all-or-none compliance, and (3) the control group receives the "standard of care".

Robins (1998) develops methodology for comparing a new therapy to the standard of care. The concern is that differences in compliance to equivalent drugs can yield results that make one drug appear more efficacious than another. Robins removes the problems of confounding by assuming that compliance is non-random and can be predicted by time-dependent prognostic factors.

Note that oral contraceptives and beta-blockers are the only drug classes with available information on how to modify dosing behavior after skipping one or more pills. The relationship between actual drug intake and response to these drugs was determined via controlled simulations of noncompliance(Morris, Groom et al. 1979; Chowdhury, Joshi et al. 1980; Wang, Shi et al. 1982; Landgren and Diczfalusy 1984; Landgren and

Csemiczky 1991; Guillebaud 1993; Johnson and Whelton 1994; Vaur, Dutrey-Dupagne et al. 1995). That is, the causal effect of noncompliance was determined by randomizing noncompliance. Certain pills in the cycle were replaced with placebo to simulate skipping pills. This causal design is one solution to the issue of confounding.

1.3.2.2 Caveat #2: Measurement Error

A second limitation of techniques for determining exposure-response using compliance data is the accuracy and precision of patient compliance measurement. Since none of the tools record the time each tablet is swallowed—arguably, the estimand of patient compliance—intake inferred from the data is prone to error.

All compliance measuring tools are subject to random and nonrandom sources of error. In self-reporting compliance, subjects may simply forget the number of pills they do not remember to take (random error), or they may intentionally inflate estimates of intake to please their care provider (nonrandom error). Pill counts have been criticized for grossly overestimating compliance(Pullar, Kumar et al. 1989). Investigator miscounting may be a source of random error in pill counts. "Pill dumping"—the act of discarding unused pills in order to appear fully compliant—is a well-documented phenomenon that yields nonrandom error in pill count compliance(Kass, Meltzer et al. 1986; Pullar, Kumar et al. 1989; Rudd, Byyny et al. 1989; Nides, Tashkin et al. 1993). Assay noise is a source of random error in biological assays of compliance, while "white coat compliance"(Feinstein 1990)—the act of improving drug taking behavior several days prior to a visit with a clinician—is a nonrandom source of error. Forgetting to take pills removed from an electronically monitored bottle is a random source of error in compliance measured using an electronic cap. Intentionally neglecting to take the removed pills is a nonrandom source of error. Note that the electronic cap requires more work on the part of the patient to yield overestimates of compliance compared to all other tools. Of all available tools, only electronic monitors are suspected to yield downwardly biased compliance estimates(Burney, Krishnan et al. 1996; Bangsberg, Hecht et al. 2000; Turner and Hecht 2001).

It is known that random error in an independent variable attenuates the estimated causal relationship with its dependent variable(Carroll 1995). That is, random error in compliance measurement yields downwardly biased estimates of the exposure-response relationship. Ironically, attenuation of the estimated drug effect relationship is the very problem with ITT that motivates the use of compliance data! Nonrandom error in the independent variable may bias the estimated drug effect relationship upward or downward. It is unknown whether attenuation is greater when assuming perfect compliance or when using a faulty measure of compliance.

The statistical literature has a long history of addressing measurement error(Carroll 1995). Correction for measurement error can be viewed as a special class of data analytic approaches within the general missing data framework. Chapter 2 provides a discussion of this work.

Experimental protocols can be altered to reduce error in compliance measurement. To decrease nonrandom error in self-reported compliance, investigators may carefully choose nonjudgmental language in eliciting compliance information(Kaplan and Simon 1990; Catania, Binson et al. 1996). Electronic diaries that time stamp entries may diminish both random and nonrandom error by reducing the reliance on patient memory and making it more difficult for patients to intentionally misrepresent their intake(Hyland, Kenyon et al. 1993). Random error in pill counts is likely negligible if investigators perform multiple counts. Unannounced pill counts—having the study investigator unexpectedly visit the subject at his place of residence to count pills—may reduce nonrandom error in pill counts as it offers the patient less of an impetus to dump pills(Bangsberg, Hecht et al. 2000). Long half-life markers can be monitored to ascertain drug intake over a longer period of time than drug concentration monitoring may allow, thus, reducing the effect of white coat compliance(Hardy, Kumar et al. 1990). Electronic measures may be corrected using self-reported compliance information(Bangsberg, Hecht et al. 2000). For example, a subject who consistently opens his pill bottle only once every day despite assignment to a b.i.d. regimen will have the electronically monitored compliance value (50%) adjusted to reflect perfect compliance (100%) if the patient reports removing doses for the entire day at one opening.

Although electronic diaries, unannounced pill counts, long half-life marker compounds, and electronically monitored caps with supplemental self-report information may provide the most accurate measure of drug intake, they are not the most common methods used in practice. Considerations of cost and convenience strongly influence the selection of compliance monitoring tools. Since compliance assessment is subject to considerable error, some recommend the use of two or more instruments(Liu, Golin et al. 2001). To satisfy the need for economy and accuracy in compliance measurement, calibration study designs—where compliance is assessed with a less accurate tool in all subjects and with a more accurate tool in a random subset—has been used in some AIDS Clinical Trials Group (ACTG) protocols.

Calibration designs bring up two important issues in addition to the measurement error problem: (1) How does one determine compliance when it is measured with several tools and the measurements do not agree, and (2) How does one use partial compliance data from one instrument in conjunction with full or partial data from another of lesser accuracy. The first issue has been addressed (with respect to drug level outcomes, not clinical outcomes)(Jonsson, Wade et al. 1997); the second has not.

1.4 References

Albert, J. M. and D. L. Demets (1994). "On a Model-Based Approach to Estimating Efficacy in Clinical Trials." Statistics in Medicine **13**: 2323-2335.

Angrist, J. D., G. W. Imbens and D. R. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables." Journal of the American Statistical Association **91**: 444-472.

Bangsberg, D. R., F. M. Hecht, E. D. Charlebois, A. R. Zolopa, M. Holodniy, L. B. Sheiner, J. D. Bamberger, M. A. Chesney and A. Moss (2000). "Adherence to Protease Inhibitors, HIV-1 Viral Load, and Development of Drug Resistance in an Indigent Population." HIV **14**(4): 357-366.

Bangsberg, D. R., F. M. Hecht, H. Clague, E. D. Charlebois, D. Ciccarone, M. Chesney and A. Moss (2001). "Provider Assessment of Adherence to HIV Antiretroviral Therapy." Journal of the Acquired Immune Deficiency Syndrome **26**(5): 435-442.

Barker, J. and J. Millard (2000). "Directly Observed Therapy and Treatment Adherence." Lancet **356**(9234): 1030-1031; discussion 1032.

Bloch, M., E. Gur and A. Y. Shalev (1992). "Hypouricemic Effect of Zuclopenthixol: A Potential Marker of Drug Compliance?" Psychopharmacology **109**(3): 377-378.

Bond, W. S. and D. A. Hussar (1991). "Detection Methods and Strategies for Improving Medication Compliance." American Journal of Hospital Pharmacy **48**(9): 1978-1988.

Brittain, E. and J. Wittes (1990). "The Run-in Period in Clinical Trials. The Effect of Misclassification on Efficiency." Controlled Clinical Trials **11**(5): 327-338.

Burney, K. D., K. Krishnan, M. T. Ruffin, D. Zhang and D. E. Brenner (1996). "Adherence to Single Daily Dose of Aspirin in a Chemoprevention Trial. An Evaluation of Self-Report and Microelectronic Monitoring." Archives of Family Medicine **5**(5): 297-300.

Carroll, R. J., Ruppert, D., and Stefanski, L.A. (1995). Measurement Error in Nonlinear Models. Great Britian, St. Edmundsbury Press.

Catania, J. A., D. Binson, J. Canchola and L. M. Pollack (1996). "Effects of Interviewer Gender, Interviewer Choice, and Item Wording on Responses to Questions Concerning Sexual Behavior." Public Opinion Quarterly **60:** 345-375.

Chesney, M. A. (2000). "Factors Affecting Adherence to Antiretroviral Therapy." Clinical Infectious Diseases. Supplement. **30**: S171-S176.

Chesney, M. A., J. Ickovics, F. M. Hecht, G. Sikipa and J. Rabkin (1999). "Adherence: A Necessity for Successful HIV Combination Therapy." HIV. Supplement. **13**: S271-S278.

Chesney, M. A., J. R. Ickovics, D. B. Chambers, A. L. Gifford, J. Neidig, B. Zwickl and A. W. Wu (2000). "Self-Reported Adherence to Antiretroviral Medications among Participants in HIV Clinical Trials: The AACTG Adherence Instruments. Patient Care Committee & Adherence Working Group of the Outcomes Committee of the Adult HIV Clinical Trials Group (AACTG)." AIDS Care **12**(3): 255-266.

Chesney, M. A., M. Morin and L. Sherr (2000). "Adherence to HIV Combination Therapy." Social Science and Medicine **50**(11): 1599-1605.

Chowdhury, V., U. M. Joshi, K. Gopalkrishna, S. Betrabet, S. Mehta and B. N. Saxena (1980). "'Escape' Ovulation in Women Due to the Missing of Low Dose Combination Oral Contraceptive Pills." Contraception **22**(3): 241-247.

Cramer, J. (1991). Overview of Methods to Measure and Enhance Patient Compliance. Patient Compliance in Medical Practice and Clinical Trials. J. A. Cramer and B. Spilker, Eds. New York, Raven Press: 3-10.

Cramer, J. A., R. H. Mattson, M. L. Prevey, R. D. Scheyer and V. L. Ouellette (1989). "How Often Is Medication Taken as Prescribed? A Novel Assessment Technique." Journal of the American Medical Association **261**(22): 3273-3277.

Cramer, J. A., R. D. Scheyer and R. H. Mattson (1990). "Compliance Declines between Clinic Visits." Archives of Internal Medicine **150**(7): 1509-1510.

Cramer, J. A. and B. Spilker, Eds. (1991). Patient Compliance in Medical Practice and Clinical Trials. New York, Raven Press Ltd.

Cummings, K. M., M. H. Becker, J. P. Kirscht and N. W. Levin (1981). "Intervention Strategies to Improve Compliance with Medical Regimens by Ambulatory Hemodialysis Patients." Journal of Behavioral Medicine **4**(1): 111-127.

De Geest, S., L. Borgermans, H. Gemoets, I. Abraham, H. Vlaminck, G. Evers and Y. Vanrenterghem (1995). "Incidence, Determinants, and Consequences of Subclinical Noncompliance with Immunosuppresive Therapy in Renal Transplant Recipients." Transplantation **59**: 340-347.

de Klerk, E., D. van der Heijde, H. van der Tempel and S. van der Linden (1999). "Development of a Questionnaire to Investigate Patient Compliance with Antirheumatic Drug Therapy." Journal of Rheumatology **26**(12): 2635-2641.

de Klerk, E. and S. van der Linden (1996). "Compliance Monitoring of NSAID Drug-Therapy in Ankylosing Spondylitis, Experiences with an Electronic Monitoring Device." British Journal of Rheumatology **35**: 60-65.

Detry, J. M. (1994). "Patient Compliance and Therapeutic Coverage: Amlodipine Versus Nifedipine Slow Release in the Treatment of Hypertension and Angina: Interim Results. Steering Committee and Cardiologists and General Practitioners Involved in the Belgium Multicentre Study on Patient Compliance." Clinical Cardiology. Supplement. **17**(9): 12-16.

Didlake, R. H., K. Dreyfus, R. H. Kerman, C. T. Van Buren and B. D. Kahan (1988). "Patient Noncompliance: A Major Cause of Late Graft Failure in Cyclosporine-Treated Renal Transplants." Transplant Proceedings. Supplement. **20**(3): 63-69.

Dunn, J., D. Golden, C. T. Van Buren, R. M. Lewis, J. Lawen and B. D. Kahan (1990). "Causes of Graft Loss Beyond Two Years in the Cyclosporine Era." Transplantation **49**(2): 349-353.

Efron, B. and D. Feldman (1991). "Compliance as an Explanatory Variable in Clinical Trials." Journal of the American Statistical Association **86**: 9-22.

Farmer, K. (1999). "Methods for Measuring and Monitoring Medication Regimen Adherence in Clinical Trials and Clinical Practice." Clinical Therapeutics **21**(6): 1074-1090.

Feinstein, A. R. (1990). "On White-Coat Effects and the Electronic Monitoring of Compliance." Archives of Internal Medicine **150**(7): 1377-1378.

Freedman, L. S. (1990). "Effect of Partial Noncompliance on the Power of a Clinical Trial." Controlled Clinical Trials **11**: 157-168.

Girard, P., T. F. Blaschke, H. Kastrissios and L. B. Sheiner (1998). "A Markov Mixed Effect Regression Model for Drug Compliance." Statistics in Medicine **17**(20): 2313-2333.

Goetghebeur, E., G. Molenberghs and J. Katz (1998). "Causal Effect of Compliance on Binary Outcome in Randomized Controlled Trials." Statistics in Medicine **17**: 341-355.

Guillebaud, J. (1993). "Any Questions." British Medical Journal **307**: 617.

Hardy, E., S. Kumar, S. Peaker, M. Feely and T. Pullar (1990). "A Comparison of a Short Half-Life Marker (Low-Dose Isoniazid), a Long Half-Life Pharmacological Indicator (Low-Dose Phenobarbitone) and Measurements of a Controlled Release

'Therapeutic Drug' (Metoprolol, Metoros) in Reflecting Incomplete Compliance by Volunteers." British Journal of Clinical Pharmacology **30**(3): 437-441.

Harter, J. G. and C. C. Peck (1991). "Chronobiology: Suggestions for Integrating It into Drug Development." Annals of the New York Academy of Sciences **618**: 563-571.

Hasford, J. (1999). Design and Analysis of Clinical Trials of Compliance. Drug Regimen Compliance: Issues in Clinical Trials and Patient Management. J. M. Metry and U. A. Meyer, Eds. Chichester, John Wiley & Sons: 23-40.

Haynes, R., Montague, P, Oliver, T, McKibbon, KA, Brouwer, MC, Kana, R. (2000). Interventions for Helping Patients to Follow Prescriptions for Medications, Cochrane Database Systematic Reviews.

Hong, J. H., N. Sumrani, V. Delaney, R. Davis, A. Dibenedetto and K. M. H. Butt (1992). "Causes of Late Renal Allograft Failure in the Ciclosporin Era." Nephron **62**: 272-278.

Hyland, M. E., C. A. Kenyon, R. Allen and P. Howarth (1993). "Diary Keeping in Asthma: Comparison of Written and Electronic Methods." British Medical Journal **306**(6876): 487-489.

Ickovic, J. R. and A. W. Meisler (1997). "Adherence in HIV Clinical Trials: A Framework for Clinical Research and Clinical Care." Journal of Clinical Epidemiology **50**(4): 385-391.

Johnson, B. F. and A. Whelton (1994). "A Study Design for Comparing the Effects of Missing Daily Doses of Antihypertensive Drugs." American Journal of Therapeutics **1**(260-267).

Jonsson, E. N., J. R. Wade, G. Almkvist and M. O. Karlson (1997). "Discrimination between Rival Dosing Histories." Pharmaceutical Research **14**: 984-991.

Kaplan, R. M. and H. J. Simon (1990). "Compliance in Medical Care: Reconsideration of Self-Predictions." Annals of Behavioral Medicine **12**(2): 66-71.

Kapur, S., R. Ganguli, R. Ulrich and U. Raghu (1991). "Use of Random-Sequence Riboflavin as a Marker of Medication Compliance in Chronic Schizophrenics." Schizophrenia Research **6**(1): 49-53.

Kass, M. A., M. Gordon and D. W. Meltzer (1986). "Can Ophthalmologists Correctly Identify Patients Defaulting from Pilocarpine Therapy?" American Journal of Ophthalmology **101**(5): 524-530.

Kass, M. A., D. Meltzer, M. Gordon, D. Cooper and J. Goldberg (1986). "Compliance with Topical Pilocarpine Treatment." American Journal of Ophthalmology **101**: 515-523.

Kastrissios, H. and T. F. Blaschke (1997). "Medication Compliance as a Feature in Drug Development." Annual Review of Pharmacology and Toxicology **37**: 451-475.

Kastrissios, H. and T. F. Blaschke (1998). "Therapeutic Implications of Nonadherence with Antiretroviral Drug Regimens." HIV **8**(2): 24-28.

Koop, C. E. (1984). Keynote Address: Improving Medication Compliance. National Pharmaceutical Council Symposium, Washington, DC.

Landgren, B. M. and G. Csemiczky (1991). "The Effect of Follicular Growth and Luteal Function of "Missing the Pill". A Comparison between a Monophasic and a Triphasic Combined Oral Contraceptive." Contraception **43**(2): 149-159.

Landgren, B. M. and E. Diczfalusy (1984). "Hormonal Consequences of Missing the Pill During the First Two Days of Three Consecutive Artificial Cycles." Contraception **29**(5): 437-446.

Lasagna, L. and P. B. Hutt (1991). Health Care, Research, and Regulatory Impact of Noncompliance. Compliance in Medical Practice and Clinical Trials. J. A. Cramer and B. Spilker, Eds. New York, Raven Press: 393-403.

Lee, J. Y., J. W. Kusek, P. G. Greene, S. Bernhard, K. Norris, D. Smith, B. Wilkening and J. T. Wright, Jr. (1996). "Assessing Medication Adherence by Pill Count and Electronic Monitoring in the African American Study of Kidney Disease and Hypertension (AASK) Pilot Study." American Journal of Hypertension **9**(8): 719-725.

Lerner, B. H., R. M. Gulick and N. N. Dubler (1998). "Rethinking Nonadherence: Historical Perspectives on Triple-Drug Therapy for HIV Disease." Annals of Internal Medicine **129**(7): 573-578.

Lipsitch, M. and B. R. Levin (1998). "Population Dynamics of Tuberculosis Treatment: Mathematical Models of the Roles of Noncompliance and Bacterial Heterogeneity in the Evolution of Drug Resistance." International Journal of Tuberculosis and Lung Disease **2**(3): 187-199.

Liu, H., C. E. Golin, L. G. Miller, R. D. Hays, K. Beck, S. Sanandaji, J. Christian, T. Maldonado, D. Duran, A. H. Kaplan and N. S. Wenger (2001). "A Comparison Study of Multiple Measures of Adherence to HIV Protease Inhibitors." Annals of Internal Medicine **134**: 968-977.

Maenpaa, H., K. Javela, J. Pikkarainen, M. Malkonen, O. P. Heinonen and V. Manninen (1987). "Minimal Doses of Digoxin: A New Marker for Compliance to Medication." European Heart Journal. Supplement. **8**: 31-37.

Meredith, P. A. (1999). Achieving and Assessing Therapeutic Coverage. Drug Regimen Compliance: Issues in Clinical Trials and Patient Management. J. M. Metry and U. A. Meyer, Eds. Chichester, John Wiley & Sons: 41-60.

Meredith, P. A. and H. L. Elliott (1994). "Therapeutic Coverage: Reducing the Risks of Partial Compliance." British Journal of Clinical Practice. Supplement. **73**: 13-17.

Mitchison, D. A. (1998). "How Drug Resistance Emerges as a Result of Poor Compliance During Short Course Chemotherapy for Tuberculosis." International Journal of Tuberculosis and Lung Disease **2**(1): 10-15.

Morris, S. E., G. V. Groom, E. D. Cameron, M. S. Buckingham, J. M. Everitt and M. Elstein (1979). "Studies on Low Dose Oral Contraceptives: Plasma Hormone Changes in Relation to Deliberate Pill ('Microgynon-30') Omission." Contraception **20**: 61-69.

Nides, M. A., D. P. Tashkin, M. S. Simmons, R. A. Wise, V. C. Li and C. S. Rand (1993). "Improving Inhaler Adherence in a Clinical Trial through the Use of the Nebulizer Chronolog." Chest **104**(2): 501-507.

Nieuwkerk, P. T., M. A. Sprangers, D. M. Burger, R. M. Hoetelmans, P. W. Hugen, S. A. Danner, M. E. van Der Ende, M. M. Schneider, G. Schrey, P. L. Meenhorst, H. G. Sprenger, R. H. Kauffmann, M. Jambroes, M. A. Chesney, F. de Wolf and J. M. Lange (2001). "Limited Patient Adherence to Highly Active Antiretroviral Therapy for

HIV-1 Infection in an Observational Cohort Study." Archives of Internal Medicine **161**(16): 1962-1968.

Pablos-Mendez, A., R. G. Barr and S. Shea (1998). "Run-in Periods in Randomized Trials: Implications for the Application of Results in Clinical Practice." Journal of the American Medical Association **279**(3): 222-225.

Peck, C. C. (1999). Non-Compliance and Clinical Trials: Regulatory Perspectives. Drug Regimen Compliance: Issues in Clinical Trials and Patient Management. J. M. Metry and U. A. Meyer, Eds. Chichester, John Wiley & Sons: 97-102.

Program, L. R. C. (1984). "The Lipid Research Clinics Coronary Primary Prevention Trial Results, Parts I and II." Journal of the American Medical Association **251**: 351-374.

Psaty, B. M., T. D. Koepsell, E. H. Wagner, J. P. LoGerfo and T. S. Inui (1990). "The Relative Risk of Incident Coronary Heart Disease Associated with Recently Stopping the Use of Beta-Blockers." Journal of the American Medical Association **263**(12): 1653-1657.

Pullar, T., S. Kumar, H. Tindall and M. Feely (1989). "Time to Stop Counting the Tablets?" Clinical Pharmacology and Therapeutics **46**(2): 163-168.

Rich, M. W., D. B. Gray, V. Beckham, C. Wittenberg and P. Luther (1996). "Effect of a Multidisciplinary Intervention on Medication Compliance in Elderly Patients with Congestive Heart Failure." The American Journal of Medicine **101**(3): 270-276.

Robins, J. M. (1998). "Correction for Non-Compliance in Equivalence Trials." Statistics in Medicine **17**: 269-302.

Rudd, P., R. L. Byyny, V. Zachary, M. E. LoVerde, C. Titus, W. D. Mitchell and G. Marshall (1989). "The Natural History of Medication Compliance in a Drug Trial: Limitations of Pill Counts." Clinical Pharmacology and Therapeutics **46**(2): 169-176.

Sheiner, L. B. and D. B. Rubin (1995). "Intention-to-Treat Analysis and the Goals of Clinical Trials." Clinical Pharmacology and Therapeutics **57**(1): 6-15.

Turner, B. J. and F. M. Hecht (2001). "Improving on a Coin Toss to Predict Patient Adherence to Medications." Annals of Internal Medicine **134**(10): 1004-1006.

Urquhart, J. (1991). Patient Compliance as an Explanatory Variable in Four Selected Cardiovascular Studies. Patient Compliance in Medical Practice and Clinical Trials. J. A. Cramer and B. Spilker, Eds. New York, Raven Press: 301-322.

Urquhart, J. (1992). "Ascertaining How Much Compliance Is Enough with Outpatient Antibiotic Regimens." Postgraduate Medical Journal. Supplement. **68**: S49-58; discussion S59.

Urquhart, J. (1994). "Partial Compliance in Cardiovascular Disease: Risk Implications." British Journal of Clinical Practice. Supplement. **73**: 2-12.

Urquhart, J. (1994). "Role of Patient Compliance in Clinical Pharmacokinetics. A Review of Recent Research." Clinical Pharmacokinetics **27**(3): 202-215.

Urquhart, J. (1997). "The Electronic Medication Event Monitor. Lessons for Pharmacotherapy." Clinical Pharmacokinetics **32**(5): 345-356.

Urquhart, J. and C. Chevalley (1988). "Impact of Unrecognized Dosing Errors on the Cost and Effectiveness of Pharmaceuticals." Drug Information Journal **22**: 363-378.

Urquhart, J. and E. De Klerk (1998). "Contending Paradigms for the Interpretation of Data on Patient Compliance with Therapeutic Drug Regimens." Statistics in Medicine **17**(3): 251-267; discussion 387-389.

Vaur, L., C. Dutrey-Dupagne and J. Boussac (1995). "Differential Effects of a Missed Dose of Trandolapril and Enalapril on Blood Pressure Control in Hypertensive Patients." Journal of Cardiovascular Pharmacology **26**: 127-131.

Volmink, J. and P. Garner (2001). Directly Observed Therapy for Treating Tuberculosis, Cochrane Database Systematic Reviews.

Vrijens, B. and E. Goetghebeur (1997). "Comparing Compliance Patterns between Randomized Treatments." Controlled Clinical Trials **18**(3): 187-203.

Waeber, B., G. Leonetti, R. Kolloch and G. T. McInnes (1999). "Compliance with Aspirin or Placebo in the Hypertension Optimal Treatment (HOT) Study." Journal of Hypertension **17**(7): 1041-1045.

Wall, T. L., J. L. Sorensen, S. L. Batki, K. L. Delucchi, J. A. London and M. A. Chesney (1995). "Adherence to Zidovudine (AZT) among HIV-Infected Methadone Patients: A Pilot Study of Supervised Therapy and Dispensing Compared to Usual Care." Drug and Alcohol Dependence **37**(3): 261-269.

Wang, E., S. Shi, S. Z. Cekan, B. M. Landgren and E. Diczfalusy (1982). "Hormonal Consequences of "Missing the Pill"." Contraception **26**(6): 545-566.

Weis, S. E., P. C. Slocum, F. X. Blais, B. King, M. Nunn, G. B. Matney, E. Gomez and B. H. Foresman (1994). "The Effect of Directly Observed Therapy on the Rates of Drug Resistance and Relapse in Tuberculosis." New England Journal of Medicine **330**(17): 1179-1184.

Wright, M. T. (2000). "The Old Problem of Adherence: Research on Treatment Adherence and Its Relevance for HIV/HIV." AIDS Care **12**(6): 703-710.

Chapter 2: Statistical Approaches to Correcting for Measurement Error and Missing Data

2.1 Error in Compliance Versus Error in Response

Scientists who work at a lab bench often consider the independent variable, X, as something that can be measured accurately, and the dependent variable, Y, as subject to measurement error. This reflects the nature of controlled experiments. In enzyme kinetic studies, for example, the experimentalist determines reactant concentration, while the product of the reaction is subject to biological variability.

In contrast, scientists who study observational data, such as survey data, often view X as an error-prone variable. Although some covariates, such as gender, can be determined accurately, many others cannot. A predictor variable such as alcohol consumption is likely to be measured with great error.

This distinction between error in X and error in Y is important to make because each has a different impact on the validity of data analysis. To understand this, first, one must be clear about what is meant by data analysis. Here, it is assumed that the goal is to develop a predictive model that quantifies the trend in Y as a function of X. For the sake of discussion, assume that the true relationship between X and Y is linear, such that

$$Y = \alpha + \beta X + \varepsilon, \tag{2.1}$$

where α and β represent the y-intercept and slope, and ε is an error term. A valid data analysis is one in which the parameter estimates, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\varepsilon}$, obtained through some model fitting procedure, are unbiased for the true values of α, β, and ε.

If equation 2.1 is fit to a data set having negligible error in X but random error in Y—the type of data commonly analyzed by bench researchers—parameter estimates will be unbiased. Thus, a standard least squares fitting procedure minimizing the distance between the model's predicted Y and measured Y (Y understood to be a vector) delivers valid model parameters when there is random error in Y. However, if equation 2.1 is fit to a data set with random error in X, parameter estimates will be biased. Noise in X acts to make the estimated slope approach zero (a phenomenon referred to as attenuation) and makes the estimated magnitude of ε larger than the true value. More disconcerting is that bias in X can cause the estimated parameters to be biased upward or downward(Carroll 1995).

Depending on how it is measured, compliance may be a biased and/or noisy covariate. Thus, a simple regression of response on exposure, where exposure is a function of compliance, is expected to yield biased parameter estimates. (Note that compliance and exposure will be referred to interchangeably as X in this discussion.) Ironically, bias in exposure-response estimation is the very problem with **ITT** motivating the use of compliance data!

Fortunately, statistical methods have been developed to obtain unbiased parameter estimates when X is measured with error. The "measurement error" literature, also known as "errors-in-variables", focuses on correcting parameter estimates obtained using standard methods of analysis for bias and/or noise in X. Measurement error can be considered a specific class of problems within the more general "missing data" framework. The approaches of these two fields are highlighted in this chapter. This general background is necessary for understanding the methods of analyses compared in this report.

2.2 Measurement Error Approaches(Carroll 1995)

In the measurement error framework, X is unobservable. The basic strategy, then, is to measure a covariate, W, that is related to X, and tease out the information it contains on X. The measurement error literature largely takes what is referred to as a "functional" modeling approach. In other words, few assumptions are made about the distribution of X. Typically, additive or multiplicative error models with error modeled as a normally distributed variable with mean zero are assumed, such as

$$W = X + \varepsilon_X.$$

Estimates of var(ε_X) can be obtained via independent replicate measures of W on subjects in the clinical trial or from an external source of data. If there exists no covariate related to X, an instrumental variable, T, can be measured instead. T is a variable that influences the response variable, Y, only via X. In this way, T is said to be "surrogate" for X, "conditionally independent of Y given X", or have "nondifferential" measurement error. To be considered an instrumental variable, T must be correlated with X, independent of W-X, and independent of Y given X. Once W or T is known, the method of "regression calibration" replaces X by the regression of X on W or X on T in the model for Y. The standard error of estimates must be corrected to account for the filled in data.

A different approach, which does not involve measuring additional covariates, is to perform a simulation extrapolation (SIMEX). The idea is simulate the effect of noise on X in order to extrapolate back to the hypothetical situation in which X has no error. To do this, increasing amounts of noise are added to X, and a model is fit to the original Y data and jittered X data. Estimated parameter values are plotted as a function of the amount of noise—some multiple of ε is plotted on the x-axis and the corresponding parameter estimate is plotted on the y-axis. The parameter value corrected for error in the measurement of X is obtained by extrapolating back to the y-intercept where $\varepsilon=0$.

A newer approach in the measurement error literature is referred to as "structural" modeling. Here, X is a random variable and a parametric model is placed on its distribution. Likelihood methods are used, meaning that one writes a model for the joint distribution of X and Y. This approach is more computationally intensive than the functional approach, although assumptions are often made to simplify the computation. One benefit is that a likelihood formulation allows computation of confidence intervals via the likelihood ratio. When nonlinear least squares is used, the confidence intervals must be determined by bootstrap or by a normal approximation.

This brings up an important point about the purpose of correcting for measurement error. If a predictor, X, is measured with error and one wants to predict a response based on some error prone measure of X, then it makes little sense to worry about the unobservable value of X. After all, Y can be predicted from W using a model in which Y is regressed on W. However, if one wants to predict responses in another individual, the relationship between X and W may not be the same in that individual as in subjects observed in the study. A naive prediction model that ignores measurement error may not be transportable. Unbiased parameter estimates are required for valid inference.

This is especially important given the challenges inherent in compliance measurement. Error in compliance measurement depends on the measurement tool available and how it is used, so the error is likely to be different during the study in which the model is created than in clinical applications. If the goal is to create dosing guidelines that are published in a package insert, the information needs to be as general as possible. Thus, it is imperative to correct for measurement error in compliance in exposure-response models.

2.3 Measurement Error Approaches are a Special Case of Missing Data Approaches

A study in which compliance is measured via a calibration design fits into the classical measurement error scheme where W (self-reported compliance) is known in all subjects and an internal validation study is used to measure X (electronically monitored compliance) in a subset. Another, perhaps less optimistic, way to view the calibration design is as a missing data problem. That is, rather than thinking of C_M data as available in a subset, C_M is considered missing in a fraction of subjects. This perspective would be more obvious had the study been designed to collect C_M in all, but data were recorded in error or some subjects didn't return their electronically monitored cap. Regardless, the data set looks the same—C_M entries are incomplete.

On the surface, this distinction appears to be a semantic issue. In the statistical literature, however, the difference is philosophical. Data analytic approaches to measurement error problems are quite different from missing data solutions. Likelihood approaches are the norm in the missing data literature, whereas "structural" modeling approaches are less common in the field of measurement error.

Classic data analytic approaches have been developed to analyze full data sets. It is, perhaps, for this reason that missing data may appear to be a more serious problem than measurement error. However, measurement error can be considered more insidious than missing data, in that it doesn't make itself known. When data are missing, one is forced to choose a data analytic approach wisely.

2.4 Missing Data Approaches(Little and Rubin 1987; Little 1992)

The default method of analysis when data are missing is to discard records from subjects who do not have the value of interest. Not only can this lead to biased estimates of treatment effects (although most often biased towards the null), but the need to contain cost and the desire to fulfill an ethical responsibility to the subjects tested motivates clinical trialists to use the available data, imperfect as it is, as efficiently as possible to learn about drug effect. Consequently, much research has focused on developing statistical methods for handling incomplete data sets.

Similar to errors-in-variables problems, the consequence of analyzing a data set with missing data "incorrectly" depends on whether X or Y is missing and what one wants to estimate. Here, the discussion is limited to cases in which X is missing when it is of interest to estimate the exposure-response model, P(Y|X).

The validity of the analysis of data sets with missing values also depends on why the data are missing (the so-called missingness mechanism). Assessing why the data are missing is something for which there is no parallel in the measurement error literature. With measurement error, it is generally held that mismeasurement reflects a limitation of the measuring device, and hence, measurement errors are random. Faulty measurement devices are only one of several reasons why data may be missing. The data may be missing by design, as when an internal validation study is performed. In the missing data literature, this is often referred to as double sampling or probability sampling. The value may be missing unintentionally, as when there is an error in recording data. Subjects may forget to answer a question or fill in a response inappropriately. Such examples of designed missingness and unintentional missingness are classified as examples of "uninformative missingness". That is, the missingness of X is independent of the value of X. In contrast, X has informative missingness if subjects are more likely to refuse to report X when X has a certain value or when subjects drop out of a study due to their value of X. Discussion of the mechanism of "missingness" is facilitated by introducing a binary indicator variable R for the missingness. R is conventionally set equal to 0 if a subject has no measure of the variable of interest and set equal to 1 if the subject has an observed value.

To relate this to the problem under consideration, one needs to consider the reason why C_M data are missing. If subjects with poor compliance discard their electronic cap or break the device to hide the fact that they haven't been taking their medication, this is an instance of informative missingness. However, if C_M is missing by virtue of planning an economical study design, missingness is uninformative. If R is independent of X (and all other variables)

$$P(R|Y,X,W) = P(R),$$

then X is said to be missing completely at random (MCAR). If the missingness of X depends on fully observed covariates and/or observed Y, X is said to be missing at random (MAR). Formally,

$$P(R|Y,X,W) = P(R|W),$$

$$P(R|Y,X,W) = P(R|Y),$$

and

$$P(R|Y,X,W) = P(R|Y,W)$$

are all instances in which X is MAR. If the missingness of X depends on the missing X value, however, then X is said to be nonignorably missing. The distinction between ignorable and nonignorable missingness is important to make because it determines what approaches will yield valid parameter estimates.

Before the use of computers in statistical research, the missing data literature focused on developing ways to get ragged data (due to missing values) to resemble a complete data set so they could be analyzed using standard approaches. Numerous approaches to filling in missing values—a process known as data imputation—were developed. Note that many of the imputation methods have the same flavor as measurement error correction approaches.

One commonly used method of imputation for longitudinal data is "last observation carried forward", where one fills in an individual's missing values with the last reported value. (The last observation carried forward approach does not apply here because this is not a longitudinal study—C_M is either measured once or not at all.) Approaches, which do apply here, include 'mean' and 'regression' imputation. In mean imputation, the mean of the set of recorded values is substituted for the missing values. In regression imputation for X, the missing values are imputed according to a model estimated by regressing Y on X in the subjects who have both. Regression imputation is analogous to regression calibration in the measurement error literature. A variant of regression imputation is stochastic regression imputation, where error is added to predicted values before filling them in. Others have substituted external data for missing data—analogous to an external validation study in the measurement error literature. Imputation can be carried out iteratively by obtaining the least square estimates with complete data, then filling in least squares estimates of all missing values and repeating the process until the optimization procedure is complete. Analogous to the measurement error approaches, it is necessary to correct standard errors for the differing status of the true and imputed data.

More recent developments, which have become the major focus of the field, are model-based procedures for handling missing data. That is, one develops a full probability model for all of the data, missing and observed, and uses likelihood or Bayes procedures. With likelihood methods, the goal is to obtain parameter estimates that make the observed data most probable given a model for all of the data—observed and unobserved. Bayesian approaches take the likelihood approach one step further and treat all parameters as random variables. In the Bayesian framework, the likelihood is multiplied by a prior distribution on all of the parameters. Multiple imputation is a method that eases the computational burden (integration) on the likelihood approach by using multiple stochastic regression imputations based on all of the observed data.

2.5 Comparison of Approaches

Given the vast choice of approaches to analyzing a calibration study, it is important to consider what is known about method performance before committing to a strategy.

The default method of analysis when data are missing—throw out records from subjects missing data and analyze the remaining data via whatever standard method was originally planned—is attractive in its ease of use. This complete data (**CD**) approach allows one to use standard statistical software for data analysis. More importantly, **CD** yields valid inference when X is nonignorably missing(Glynn and Laird 1986; Little 1992). The disadvantage of this method is that it is inefficient.

Single imputation (SI) methods are also appealing because they allow standard methods of analysis to be used. However, these fill-in approaches are invalid if the data are not MCAR. So, if C_M data are more likely to be missing in subjects who have a worse prognosis or in subjects who have a particular value of C_Q, then SI is expected to yield biased estimates of exposure-response. Furthermore, these approaches require, oftentimes, ad hoc procedures for correcting the standard errors to reflect the differing status of real data and filled in data. While it is easy to correct residual error and standard errors that have one degree of freedom, it becomes increasingly difficult as the number of degrees of freedom increases. This is also an issue for the regression calibration approaches in the measurement error literature.

Model-based methods have gained favor over early imputation procedures for many reasons. Under the assumed model, large sample properties (consistency, efficiency) hold for likelihood methods. The likelihood ratio is asymptotically distributed chi-square—a favorable property for inference. Maximum likelihood (**ML**) procedures yield valid parameter estimates under a less stringent condition of missingness—parameter estimates are unbiased when the data are MAR. (Recall that the imputation methods require the data to be MCAR.) Furthermore, likelihood estimates can be made valid even if the data are informatively missing by modeling P(R|X), although the validity cannot be tested on the data at hand.

Under certain conditions, estimates by imputation procedures approach ML estimates. Single imputation followed by weighted least squares estimation has been shown to yield parameter estimates that approach ML estimates when the weights are chosen appropriately. Parameter estimates obtained via multiple imputation converge on likelihood estimates as the sample size increases and the number of imputations increases.

There is one instance in which data imputation followed by weighted least squares estimation performs better than likelihood methods. Elliptical distributions with a long tail are better estimated by weighted least squares procedures if data are nonnormally distributed and ML assumes normality. This highlights the fact that model based methods may be sensitive to model misspecification. Some perceive the need to write a model for all of the data and the resulting requirement to investigate sensitivity to assumptions as disadvantages of model-based methods.

Both likelihood and least squares approaches are poor for small sample inference. Bayesian (BY) approaches perform well under small sample designs. Some view it as a disadvantage that BY requires more modeling assumptions than ML. Additionally, BY is more computationally taxing.

2.6 How Should Compliance Calibration Studies Be Analyzed?

That the analysis of calibration studies has received much attention in the literature is reflected in the wide array of terminology used to describe this design. A literature search for methodology should include keywords such as "two-phase sampling", "two-level data", "coarse classification", "detailed and crude sampling", and analysis using "auxiliary data" on an "incomplete", "missing", or "mismeasured" covariate. One might assume, then, that practical guidelines exist on how to determine the exposure-response relationship using compliance data measured via a calibration study. Unfortunately, they do not.

The published explorations have been tailored to individual pockets of research with study designs, parameters of interest, and performance metrics differing from our concerns(Weinkam, Rosenbaum et al. 1991; Plummer and Clayton 1993; Bashir and Duffy 1997; Lu, Ye et al. 1997; Spiegelman, Schneeweiss et al. 1997; Golm 1998). It is easier to apply results generally between one study and another when linear models are used. However, clinical trialists are generally interested in nonlinear and even nonparametric exposure-response models. A model estimating relative risk of some outcome raises different concerns than an E_{max} model.

Perhaps, the most important difference is that reports in the literature tend to focus on large studies, typical of epidemiological research, rather than those on the order of typical confirmatory clinical trials(Rosner, Willett et al. 1989; Carroll, Freedman et al. 1997; Kaaks and Riboli 1997; Thoresen and Laake 2000). This is an important distinction, as it is known that model-based methods perform differently under asymptotic conditions compared with small sample conditions. Of course, the cutoff for what can be considered a small study with regard to asymptotic behavior is not known, but, most likely, one AIDS Clinical Trials Group data set that we aim to analyze is a small sample study—there are 34 subjects in the trial with even fewer having an accurate measure of intake(Bangsberg, Hecht et al. 2000). More details on this data set are given in Chapter 7. One study comparing regression imputation to likelihood methods for the analysis of categorical data with sample sizes of the magnitude of interest here did not address the kinds of parameters relevant to these studies(Selen 1986).

Additionally, former investigations into calibration methodology occurred in fields that use observational data. Therefore, the focus is on picking a best method among missing data or measurement error approaches. In the realm of confirmatory clinical trial analysis, the standard of comparison is the intention-to-treat procedure. An important question that needs to be explored is whether parameters in an exposure-response model suffer greater bias from using compliance data incorrectly or from discarding it altogether.

2.7 References

Bangsberg, D. R., F. M. Hecht, E. D. Charlebois, A. R. Zolopa, M. Holodniy, L. B. Sheiner, J. D. Bamberger, M. A. Chesney and A. Moss (2000). “Adherence to Protease Inhibitors, HIV-1 Viral Load, and Development of Drug Resistance in an Indigent Population.” AIDS **14**(4): 357-366.

Bashir, S. A. and S. W. Duffy (1997). “The Correction of Risk Estimates for Measurement Error.” Annals of Epidemiology **7**: 154-164.

Carroll, R. J., L. Freedman and D. Pee (1997). “Design Aspects of Calibration Studies in Nutrition, with Analysis of Missing Data in Linear Measurement Error Models.” Biometrics **53**: 1440-1457.

Carroll, R. J., Ruppert, D., and Stefanski, L.A. (1995). Measurement Error in Nonlinear Models. Great Britian, St. Edmundsbury Press.

Glynn, R. J. and N. M. Laird (1986). Regression Estimates and Missing Data: Complete-Case Analysis. Boston, Harvard School of Public Health, Department of Biostatistics.

Golm, G. T., Halloran, M.E., Longini Jr., I.M. (1998). "Semi-parametric Models for Mismeasured Exposure Information in Vaccine Trials." Statistics in Medicine **17**: 2335-2352.

Kaaks, R. and E. Riboli (1997). "Validation and Calibration of Dietary Intake Measurements in the EPIC Project: Methodological Considerations." International Journal of Epidemiology **26**(1): S15-S25.

Little, R. J. A. (1992). "Regression with Missing X's: A Review." Journal of the American Statistical Association **87**: 1227-1237.

Little, R. J. A. and D. B. Rubin (1987). Statistical Analysis with Missing Data. New York, John Wiley & Sons.

Lu, Y., K. Ye, A. K. Mathur, S. Hui, T. P. Fuerst and H. K. Genant (1997). "Comparative Calibration without a Gold Standard." Statistics in Medicine **16**: 1889-1905.

Plummer, M. and D. Clayton (1993). "Measurement Error in Dietary Assessment: An Investigation Using Covariance Structure Models. Part II." Statistics in Medicine **12**: 937-948.

Rosner, B., W. C. Willett and D. Spiegelman (1989). "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Systematic within-Person Measurement Error." Statistics in Medicine **8**: 1051-1069.

Selen, J. (1986). "Adjusting for Errors in Classification and Measurement in the Analysis of Partly and Purely Categorical Data." Journal of the American Statistical Association **81**: 75-81.

Spiegelman, D., S. Schneeweiss and A. McDermott (1997). "Measurement Error Correction for Logistic Regression Models with an "Alloyed Gold Standard"." American Journal of Epidemiology **145**(2): 184-196.

Thoresen, M. and P. Laake (2000). "A Simulation Study of Measurement Error Correction Methods in Logistic Regression." Biometrics **56**: 868-872.

Weinkam, J. J., W. L. Rosenbaum and T. D. Sterling (1991). "A Practical Approach to Estimating the True Effect of Exposure Despite Imprecise Exposure Classification." American Journal of Industrial Medicine **19**: 587-601.

Chapter 3: Scope of the Investigation

The goal of this investigation is to determine the operating characteristics of various methods for estimating "exposure-response" given a clinical trial designed such that exposure (compliance) is measured via a calibration study. To be explicit about the calibration design—compliance is measured with a biased tool in all subjects, but with an accurate tool in a random subset. Under the calibration study design there are several contending methods for determining drug exposure. For example, one may opt to only use data from subjects with an accurate measure of compliance. Alternatively, one could use all subjects' data and pick the best compliance data available in each individual to determine exposure. The purpose of this investigation is to recommend one or another analysis method for future use by others.

In this thesis, the performance of an analysis method that calibrates the imprecise measure of compliance to an accurate measure in the determination of exposure is compared to analysis methods used in practice. Methods in practice include using one, or both, or neither (neither = Intention to Treat assumes C=100% in all) of electronically monitored and self-reported compliance data as measured (not calibrated). A maximum likelihood (ML) procedure is selected as the calibration approach because it uses methodology that is acceptable in both the missing data and measurement error literature and various estimators converge on ML estimates. The missing data literature suggests that ML has difficulty under nonasymptotic conditions and that a Bayesian approach is better suited to small studies. Here, this handicap on ML performance is viewed as a way to allow the contending methods to compete with the model-based approach.

Analysis method performance is evaluated through simulation—a three-part procedure. First, data are simulated from known models of dose assignment, compliance, and exposure-response. Next, simulated data are analyzed to yield an estimate of drug effect by all methods under consideration. Finally, the precision of each drug effect estimate is computed. The Appendix outlines the simulation study design, data simulation models, and performance evaluation. Chapter 4 presents the theoretical development of data analytic methods. Chapters 5, 6, and 7 present results. Refer to the ABBREVIATIONS AND SYMBOLS section for an explanation of the notation used throughout this report and to the Appendix for a description about how to interpret the graphical displays of results.

Chapter 5 presents the investigation of analysis method performance under a best-case scenario. That is, when all assumptions of the ML method are satisfied in data simulation. Chapter 5 specifically answers:

- In general, across widely varying clinical trial designs, which method possesses the greatest estimation precision?
- Across widely varying clinical trial designs, what benefit (if any) is reaped when compliance data are rich (here, "rich" means that self-reported compliance is accurate or many subjects have C_M data)?
- Across widely varying clinical trial designs, what cost (if any) is incurred when compliance data are poor (here, "poor" means that self-reported compliance has no correlation with true intake or few subjects have C_M data)?
- To what extent does method performance depend on the size of the drug effect?
- What is the influence of the distribution of true compliance on performance?

- Which method has the most power to detect a drug effect? Does this agree with performance with respect to estimation precision?

Chapter 5 also determines the influence of arbitrary simulation study design choices on method performance. Specifically addressed—how does the prior value on $P(C|C_Q)$ and the placebo-controlled design influence method performance? Chapter 4 explains the use of the prior in estimation. The clinical trial design is outlined in the Appendix.

Chapter 6 presents a sensitivity analysis, or an investigation of method performance under conditions in which data simulation violates assumptions of the **ML** method. Refer to the "Simulation Study Design" section of the Appendix for a discussion of the assumptions of data simulation and refer to Chapter 4 for a discussion of the assumptions of the **ML** method of analysis. Chapter 6 specifically answers questions pertaining to compliance measurement:

- Across widely varying clinical trial designs, what is the influence of the assumption that the electronic monitor measures true compliance (***A3***) on method performance?
- Across widely varying clinical trial designs, what is the influence of the assumption of random assignment to electronic monitoring (***A2***) on method performance?
- Across widely varying clinical trial designs, what is the simultaneous influence of ***A2*** and ***A3*** on method performance?
- How sensitive is method performance to the source of parameter values for $P(C_Q|C)$ and $P(Y|D)$? This is investigated by asking how method performance evaluated using real data as the source for $P(Y|D)$ and $P(C_Q|C)$ in simulation

compares to performance determined using arbitrary P(Y|D) and P(C_Q|C) distributions?

Chapter 7 presents an analysis of a clinical data set using the methods investigated. Chapter 8 critically evaluates the methodology used and summarizes what has been learned in this investigation. Chapter 9 explains the impact of this work and poses future directions for this project.

SECTION II

INVESTIGATION OF DATA ANALYSIS METHODS THAT ACCOUNT FOR COMPLIANCE IN DETERMINING DRUG EXPOSURE-RESPONSE

Chapter 4: Theoretical Development of Analysis Methods

4.1 Assumptions

A1 Assignment to D_n is random

A2 Assignment to M is random

A3 The electronic monitor measures true compliance

$$C = C_M, \text{ or } P(C,C_M) = \begin{cases} 1 \text{ if } C=C_M \\ 0 \text{ otherwise} \end{cases}$$

A4 True compliance, electronically monitored compliance, and self-reported compliance do not confound P(Y|D), formally:

$P(Y|D,C) = P(Y|D)$; Y and C are conditionally independent given D

$P(Y|D,C_M) = P(Y|D)$; Y and C_M are conditionally independent given D

$P(Y|D,C_Q) = P(Y|D)$; Y and C_Q are conditionally independent given D

A5 C is a baseline covariate

4.2 Theory

In this thesis, a likelihood method for calibrating C_Q to C_M in estimating exposure-response is compared to approaches that use compliance data as measured. Likelihood methods require writing a model for all of the data, thus, we begin by writing a model for the joint distribution of all variables.

Since the individual-specific covariates, C_{Qi}, D_{ni}, and M_i, are known in all subjects, and their distributions are of no intrinsic interest, we may consider only models conditional upon them. With this conditioning, individuals are considered independent. The likelihood for all the data can therefore be written

$$L(Y_1,Y_2,\ldots Y_N,C_1,C_2,\ldots,C_N|\, C_{Q1},C_{Q2},\ldots,C_{QN},D_{n1},D_{n2},\ldots,D_{nN},M_1,M_2,\ldots,M_N) = \prod_{i=1,n} l_i,$$

where

$$l_i = P(Y_i,C_i|C_{Qi},D_{ni},M_i).$$

We now consider $l_i = P(Y_i,C_i|C_{Qi},D_{ni},M_i)$ further, omitting the individual subscript for convenience. *A2* allows one to drop the conditioning on M, and standard probability factorization allows:

$$P(Y,C|C_Q,D_n) = P(Y|C,C_Q,D_n)\ P(C|C_Q,D_n).$$

A4 allows one to rewrite the conditioning of Y on C, C_Q, and D_n as an effect of D, and *A1* and *A5* allow one to drop the conditioning of C on D_n:

$$P(Y,C|C_Q,D_n) = P(Y|D)\ P(C|C_Q) \qquad (4.1)$$

Our goal is to estimate the pharmacodynamic model parameters, θ_1. If θ_1 and θ_2 were distinct and C was known, then one might choose to model only Y|D. In this case, the second factor in the complete (individual) data likelihood (4.1) could be ignored, as it contains no information on θ_1. If, however, C is missing, as it is by design in subjects with M=0, these data may contribute to the estimation of θ_1 by integrating (4.1) over the missing C data. That is,

$$P(Y|C_Q,D_n) = \int P(Y|D)\ P(C|C_Q)dC.$$

The individual likelihood, allowing for the possibility of missing C data, is then:

$$l_i = M \bullet P(Y,C|C_Q,D_n) + (1\text{-}M) \bullet P(Y|C_Q,D_n),$$

which, by *A3*, is equivalent to:

$$l_i = M \bullet P(Y|D(C_M,D_n))P(C_M|C_Q) + (1\text{-}M) \bullet \int P(Y|\, D(C_M,D_n))P(C_M|C_Q)dC_M \quad (4.2)$$

and is now a function only of the observed data. The variables Y, D_n, C, C_M, C_Q, and D assume only a finite set of fixed values. Therefore, the likelihood (4.2) involves only summation, not integration:

$$l_i = M \bullet P(Y|D(C_M,D_n))P(C_M|C_Q) + (1\text{-}M) \bullet \sum_{C_M} P(Y|\, D(C_M,D_n))P(C_M|C_Q) \quad (4.3)$$

A consequence of the saturated parameterization is that there is an element of θ_1 for every unique combination of values of Y and D, and an element of θ_2 for every unique combination of values of C and C_Q. The element of θ_1 corresponding to a given D value is unidentifiable if no individual in a trial has that D value. For all elements of θ_2 to be identifiable, at least one subject with each possible value of C_Q must be observed in the M=1 group. Chance violation of these restrictions is likely in studies with small N. Rather than limiting this investigation to large studies, identifiability is assured (and, consequently, the task of comparing method performance is simplified) by incorporating into the likelihood a quadratic penalty for deviations of θ from a fixed prior value. Equation 4.4 formalizes this contribution.

$$L(\theta) = \pi(\theta)\prod_{i=1,n} l_i, \quad (4.4)$$

where l_i is given by (4.3), and $\pi(\theta) = \pi_1(\theta_1)\pi_2(\theta_2)$ is the penalty factor, defined as follows

$\pi_1(\theta_1) = \lambda_1(t(\theta_1) - t(\theta_{1prior}))^2$, and

$\pi_2(\theta_2) = \lambda_2(t(\theta_2) - t(\theta_{2prior}))^2$.

Note that **t** is a transformation from the unit (probability) interval to the real line, θ_{1prior} = prior mean value of θ_1, θ_{2prior} = prior mean value of θ_2, and λ_1 and λ_2 are the weights attached to the penalties (chosen arbitrarily). This corresponds to a normal prior on $t(\theta_1)$, $t(\theta_2)$. Inclusion of the penalty terms enables all analysis methods to return estimates for all parameters regardless of available data. The fewer data available to speak to a parameter, the closer its estimate will be to its prior mean.

4.3 Analysis Methods

A natural estimator of θ is the **ML** ("Maximum (penalized) Likelihood") estimator θ^{ML}; where $\theta^{ML} = \text{argmax}_\theta L(\theta)$. As noted previously, the summation in (4.3) could be avoided, as could estimation of θ_2, if C were known. This is the basis of a number of simpler pseudo-likelihood methods of estimation, which are compared to the **ML** estimator.

The first of these, denoted the **BSR ("Believe Self Report") estimator**, assumes $C = C_Q$, and estimates θ_1 as $\theta_1^{BSR} = \text{argmax}_\theta L_{BSR}(\theta)$, where

$$L_{BSR}(\theta) = \pi_1(\theta_1)\prod_{i=1,n} P(Y_i|D(C_{Qi},D_{ni})).$$

A more sophisticated variant of **BSR** is to substitute $P(Y_i|D(C_{Mi},D_{ni}))$ for $P(Y_i|D(C_{Qi},D_{ni}))$ when $M_i = 1$. This is denoted the **BA ("Best Available") estimator**, yielding θ_1^{BA}, maximizing

$$L_{BA}(\theta) = \pi_1(\theta_1)\prod_{\{i:Mi=1\}} P(Y_i|D(C_{Mi},D_{ni})) \prod_{\{i:Mi=0\}} P(Y_i|D(C_{Qi},D_{ni})).$$

On the other hand, when M=1 data are plentiful, rather than use the possibly biased C_Q at all, one might choose to use the **CD ("Complete Data")** estimator, yielding θ_1^{CD}, maximizing

$$L_{CD}(\theta)=\pi_1(\theta_1)\prod_{\{i:Mi=1\}} P(Y_i|D(C_{Mi},D_{ni})).$$

Of course, not only C_Q, but, all compliance information can be discarded (an attractive idea if compliance is suspected to be a confounder or is measured with great error), leaving one with the usual **ITT ("Intention To Treat") estimator** which regards all C = highest possible value, and yields θ_1^{ITT}, maximizing

$$L_{ITT}(\theta)=\pi_1(\theta_1)\prod_{i=1,n}P(Y_i|D=D_{ni}).$$

The theoretically most precise estimator, computed as a fiducial point for the others, is the **ALL ("All subjects have a measure of C_M") estimator**. θ_1^{ALL} maximizes

$$L_{ALL}(\theta) = \pi_1(\theta_1)\prod_{i=1,n}P(Y_i|D(C_{Mi},D_{ni})).$$

Of course, the **ALL** estimator is not a real option for any study with $f_M<1$.

Chapter 5: Investigation of Method Performance Under Ideal Conditions

Abstract

The results of chapter 5 show that **ML** yields the most precise estimates of exposure-response over widely varying clinical trial designs, extremes in quality of compliance information, and a range of drug effect sizes when data are simulated under a model in which the electronic monitor measures true compliance (formally: C_M=C) and the relationship between self-reported compliance and true compliance is not influenced by the presence of an electronic monitor (formally: $P(C_Q|C,M)=P(C_Q|C)$). **ML** is most advantageous for analyzing data sets with sparse compliance information. That is, when less than half of the subjects in a trial have C_M data and when there is no correlation between C_Q and C. Under simulation conditions selected to favor the power of one naïve analysis method over all others, **ML** is consistently the second most powerful method of analysis.

5.1 Purpose

The aim of this set of investigations is to explore method behavior under ideal conditions. That is, when data arise from a model in which assumptions of the **ML** method are satisfied (refer to Chapter 4). These results are expected to show **ML** in the best light. If there is no advantage of **ML** in this case, there is certain to be none when its assumptions are violated.

Performance of **ML**, **BSR**, **BA**, **CD** and **ITT** (and **ALL**) is investigated as a function of clinical trial size, fraction of subjects electronically monitored, accuracy of self-reported compliance, and drug effect size by analyzing data simulated from a range of N, f_M, a, b, and ρ values. Different regions of parameter space are explored to probe different aspects of method behavior. The bound on each parameter's range varies between studies, depending on the question of interest.

Refer to the Appendix for general information on the simulation study design, performance evaluation, and a key to interpretation of all graphical displays of results.

5.2 Parameters Common to All Ideal Condition Studies

Unless otherwise noted, the following parameter values (see Table 5.1) are common to all studies investigating method performance under ideal conditions.

Table 5.1. Fixed Parameter Values Common to All Investigations of Performance Under Ideal Conditions.

Parameter	**Fixed Value**
N_{lhs}	100 ×5
P(C)	P(C=0)=P(C=.5)=P(C=1) = 1/3
λ_1	.5/3
λ_2	.5/6
θ_{1prior}	$P(Y\|D=0)=P(Y\|D=.5D_n)=P(Y\|D=D_n) = .5$
θ_{2prior}	$P(C=k\|C_Q=j)=1/3$ for j,k {0,.5,1}
$P(C_M\|C)$	$C_{M,M=C}$ (see Figure A.5)

5.3 Investigation of Performance Over a Wide Range of Parameter Space—Study 1

First, it is desirable to investigate method performance across a wide range of parameter values. Table 5.2 summarizes the parameter values explored.

Table 5.2. Distribution of Random Parameter Values in Study 1.

Parameter	**Distribution**
N	U(50,400)
f_M	U(0,1)
a, b	U(0,1), U(0,2)
ρ	U(logit(.1),logit(.5)); i.e. P(Y|D=0) ranges from .1-.5

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 5.3. Note that cor(C_Q,C) is computed on simulated C_Q,C data. The range of cor(C_Q,C) values is not uniformly distributed, however, as only the microconstants comprising P(C_Q|C) (*a*,*b*) are drawn from uniform distributions. (Refer to Figure 8.1 for the histogram of cor(C_Q,C).)

The results of this study indicate that **ML** yields the greatest estimation precision across a wide range of conditions. Figure 5.1, a boxplot of the study error, Δ_R, (defined in the Performance Evaluation section of the Appendix) for study 1 shows that **ML** yields the most precise estimates of P(Y|D). (Δ_R for **ML** is closest to 1, hence, closest to **ALL**).

Table 5.3. Observed Range of Parameter Values in Study 1.

Parameter	Range
N	50—400
f_M	.00273—.997
$cor(C_Q,C)$	.0546—.801
$P(Y\|D=0)$	.100—.500
$P(Y\|D=D_n)$	.500—.900
P(C)	P(C=0): .2—.5, P(C=.5): .1875—.441, P(C=1): .154—.471

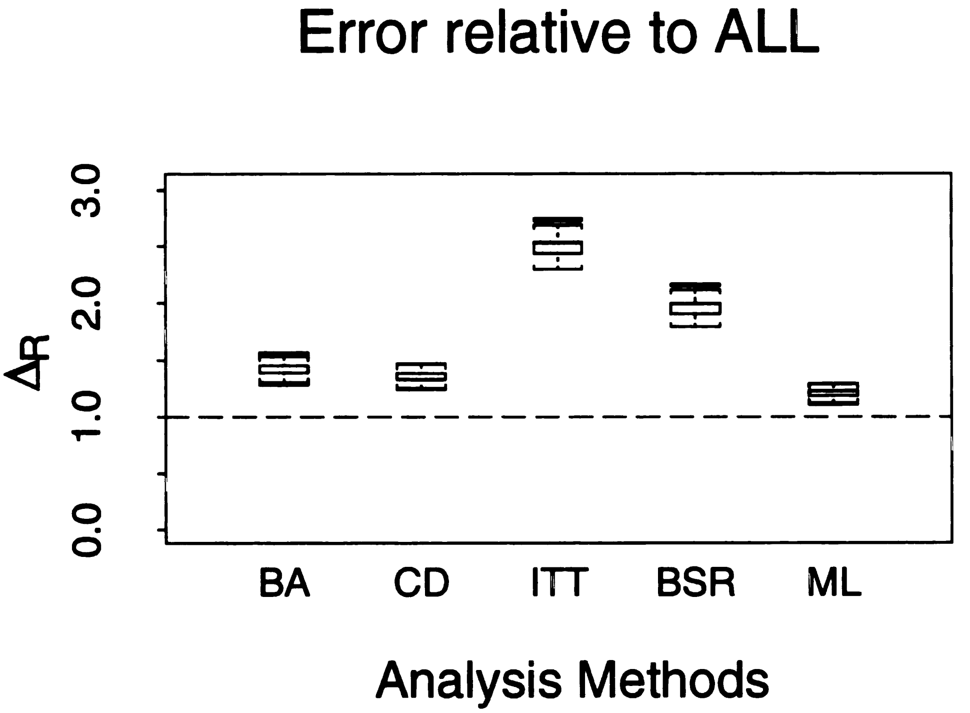


Figure 5.1. Boxplot of Δ_R for Study 1. Relative to **ALL** (dotted line), **ML** returns the most precise estimates of drug effect in a study consisting of 500 simulated clinical trials with parameters N, f_M, $cor(C_Q,C)$, and P(Y|D) ranging as in Table 5.3.

Conversely, **ITT** has the least precise and most biased estimates of P(Y|D). **BSR**, **BA**, and **CD** drug effect estimates distribute between these two extremes.

Practically speaking, this result lends support to the use of either **ML** or **CD**. That is, **ML** may yield more precise estimates of drug effect than **CD**, but **CD** is less computationally "expensive" than **ML**. Since the definition of computationally expensive differs among data analysts, methods will only be compared, here, in terms of estimation precision. The determination of merit with respect to more fluid measures of performance is considered in Chapter 8: General Findings and Recommendations. Hence, given no other information about the study design, except that the parameters fall within the wide ranges explored in study 1, Figure 5.1 suggests that it is best to use the **ML** method.

Since the Δ_R statistic distills method performance down to one number, it is conceivable that other analysis methods may outperform **ML** in specific regions of parameter space. Plots of error in estimates of P(Y|D=D_n), ($\delta_{D=Dn}$ is defined in the "Performance Evaluation" section of the Appendix) as a function of simulation parameter values reject this possibility. Figures 5.2-5.4 show that **ML** outperforms all other methods in every region of parameter space explored. (Note that plots of $\delta_{D=0}$ as a function of parameter values are not shown because the result is trivial—there is high precision in P(Y|D=0) estimates for all analysis methods. Since half of all subjects are assigned to D_n=0, and, consequently, have D=0, all methods have an abundance of valid data with which to estimate P(Y|D=0).)

Figure 5.2, a plot of $\delta_{D=Dn}$ as a function of cor(C_Q,C), shows that **ML** estimates P(Y|D=D_n) with greater precision than any other method for all values of cor(C_Q,C). Interestingly, like **CD** and **ITT**—methods that do not use C_Q data—the precision of

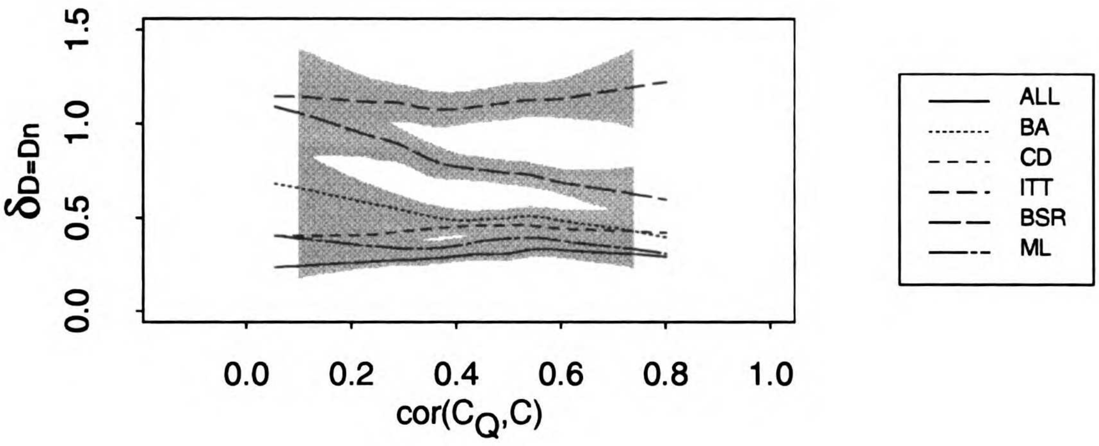


Figure 5.2. Plot of $\delta_{D=Dn}$ as a Function of cor(C_Q,C) for Study 1. ML is the only method of those using C_Q data that is insensitive to cor(C_Q,C)—the precision of its P(Y|D=D_n) estimates is equivalent for trials with cor(C_Q,C) ranging from .1 to .7.

ML's drug effect estimates are equivalent for all values of cor(C_Q,C). In contrast, all other methods that use C_Q data—**BA** and **BSR**—yield worse estimates of drug effect as cor(C_Q,C) decreases. **ML**'s insensitivity to cor(C_Q,C) is a beneficial property as the relationship between self-reported compliance and true compliance is unknown and may vary widely depending on the method used to elicit a patient's C_Q.

Figure 5.3, a plot of $\delta_{D=Dn}$ as a function of f_M, shows that **ML** yields the most precise estimates of drug effect for all values of f_M. The precision of all estimators of P(Y|D=D_n) that use C_M data eventually converges on **ML**. **CD**'s precision approaches that of **ML** at $f_M \geq .8$ and **BA** does so at $f_M \geq .9$. **ML** yields the most benefit when f_M is less than .5. In summary, there is no cost of performing the **ML** analysis for f_M greater than .5, yet there is a potential for gain when C_M data are available in fewer than half of the subjects in the trial. This is beneficial, as an investigator may not know exactly how

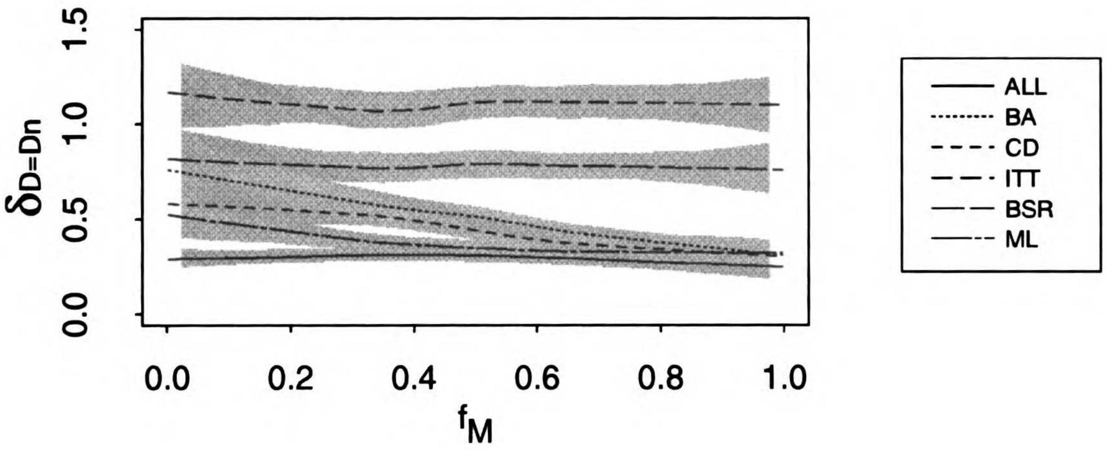


Figure 5.3. Plot of $\delta_{D=Dn}$ as a Function of f_M for Study 1. ML yields better estimates of $P(Y|D=D_n)$ than any other method (excluding **ALL**) when f_M <.8, at which point **CD** is as precise as **ML.**

much C_M data will be missing at the end of a trial when committing to a data analytic approach at the outset given that some subjects may not return electronic caps. Choosing a data analytic approach after looking at the data is often frowned upon.

Figure 5.4, a plot of $\delta_{D=Dn}$ as a function of ρ, elucidates another important property of **ML**. It is insensitive to drug effect size. In contrast, the precision of **BA**, **BSR**, and, especially, **ITT** estimates of drug effect vary with ρ. The relationship between $\delta_{D=Dn}$ and ρ indicates the extent to which a method misclassifies patient drug exposure. The larger the true difference in effect between patients receiving D=0 and $D=D_n$, the larger impact misclassifying exposure will have on estimation of $P(Y|D)$. Taken to the limit, when there is no drug effect, or ρ=0, misclassification of exposure has no influence on estimates of exposure-response since response is independent of exposure. Dependency on ρ is particularly problematic as drug effect size is not an investigator

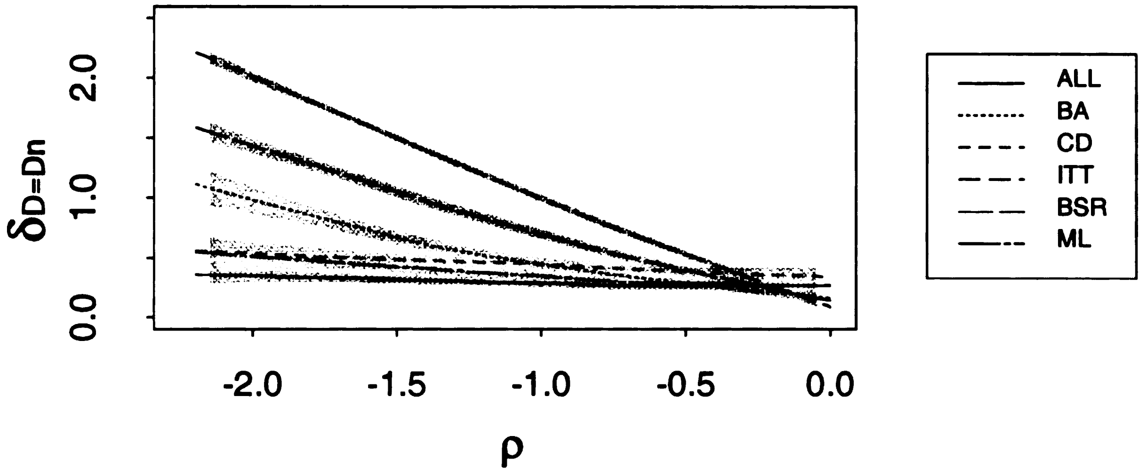


Figure 5.4. Plot of $\delta_{D=Dn}$ as a Function of ρ for Study 1. ML estimates of $P(Y|D=D_n)$ as a function of ρ are the most precise of all methods. This is a desirable property, since drug effect size is not under the investigator's control.

controlled study variable.

5.4 Investigations of Performance at Specific Locations in Parameter Space

Much is revealed about a method's performance by focusing on its behavior in extreme regions of parameter space. Of specific interest is performance when self-reported compliance is entirely accurate or, at the opposite end of the spectrum, when C_Q has no correlation with C. Additionally, it is of interest to determine method performance when electronically monitored compliance data are rich or sparse. In this discussion, "rich" is used to describe an abundance of C_M data or C_Q data that has a high correlation with C. "Poor" is used to describe sparse C_M data or C_Q data that are poorly correlated with C.

Note that some information about method performance is available at the limits in the plots of $\delta_{D=Dn}$ vs cor(C_Q,C) and $\delta_{D=Dn}$ vs f_M for study 1 (i.e. observe method performance at the lowest and highest values on the x-axis of Figure 5.2 and Figure 5.3). However, conclusions drawn in this way are unreliable since the smoothing procedure may distort performance at the edge of the plots. This is particularly problematic for Figure 5.2. Since cor(C_Q,C) is not selected by Latin Hypercube Sampling, points are not evenly distributed across the figure's x-axis. There are fewer samples at the extreme values of cor(C_Q,C) than for interior points since the P(C_Q|C) model can only yield cor(C_Q,C)=0 or cor(C_Q,C)=1 if $WT_{Q\perp C}$=1 or $WT_{Q=C}$=1, respectively. Thus, a more focused investigation is warranted.

5.4.1 Investigation of Extremes in the Quality of Self-Reported Compliance

This study design investigates the performance of methods given two extremes in the accuracy of self-reported compliance. First, the performance when all patients report compliance correctly, or cor(C_Q,C)=1, is considered (rich C_Q data). In the second simulation study, there is little correlation between a patient's self-reported compliance and his true intake (poor C_Q data). For the sake of comparison, in these studies, all parameters but those determining P(C_Q|C) range as in study 1. Refer to Table 5.4 for a summary of parameter values used in simulation. Boldface type is used to indicate parameter values drawn from ranges differing from study 1 (compare Table 5.4 to Table 5.2).

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 5.5.

Table 5.4. Distribution of Random Parameter Values in Study 2 and Study 3.

Parameter	**Distribution (or Fixed Value)**
N	U(50,400)
f_M	U(0,1)
a*, *b	**Rather than varying *a* and *b*, P(C_Q|C) is constructed by**
	directly setting $\mathrm{WT}_{Q=C}$, $\mathrm{WT}_{Q\perp C}$, and $\mathrm{WT}_{Q\geq C}$.
	Study 2: $\mathrm{WT}_{Q=C}$=1, $\mathrm{WT}_{Q\perp C}$=0, and $\mathrm{WT}_{Q\geq C}$=0
	Study 3: $\mathrm{WT}_{Q=C}$=0, $\mathrm{WT}_{Q\perp C}$=1, and $\mathrm{WT}_{Q\geq C}$=0
ρ	U(logit(.1),logit(.5))

Table 5.5. Observed Range of Parameter Values in Study 2 and Study 3.

Parameter	**Range**
N	51—400
f_M	0—1
cor(C_Q,C)	**Study 2: 1**
	Study 3: -.165—.307
P(Y|D=0)	.100—.499
P(Y|D=D_n)	.501—.900
P(C)	P(C=0): .173—.472, P(C=.5): .208—.558, P(C=1): .182—.435

5.4.1.1 Rich C_Q Data: Self-Reported Compliance is Accurate—Study 2

Figure 5.5, a boxplot of Δ_R for study 2, shows that **BA** and **BSR** yield the most precise estimates of exposure-response. Since C_Q=C, they are equivalent to **ALL**. **ML** performs nearly as well as **BA** and **BSR**, and is more precise than in study 1 (compare Figure 5.5 to Figure 5.1). The precision of **CD** and **ITT** is unchanged compared to study 1 since neither uses C_Q data.

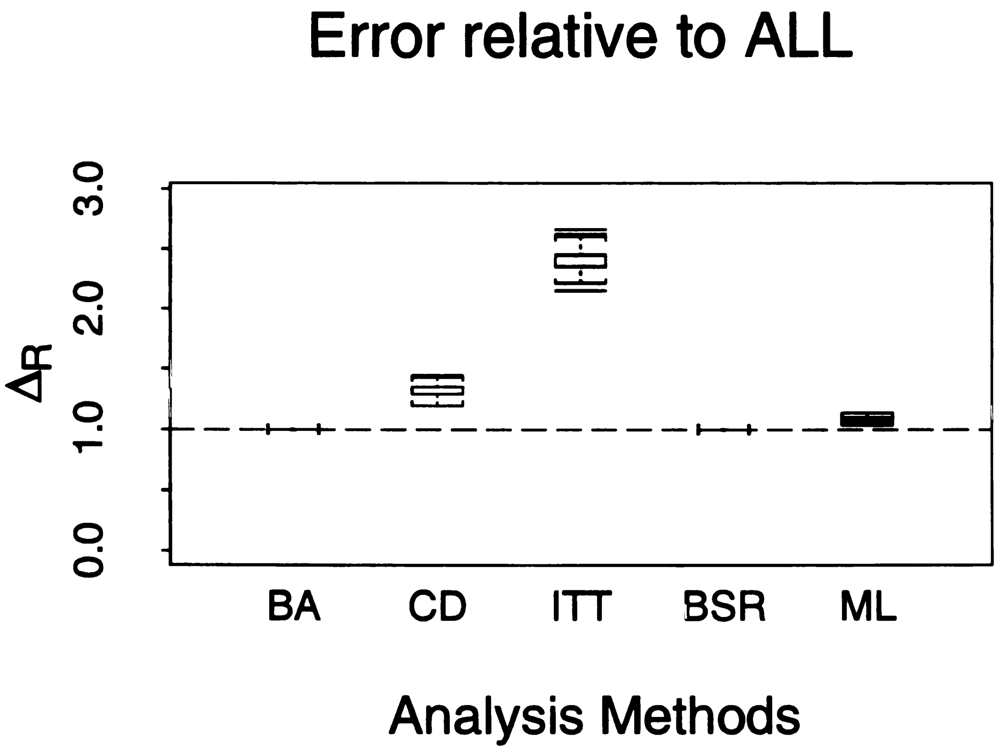


Figure 5.5. Boxplot of Δ_R for Study 2. When patients accurately self-report their intake, **BA** and **BSR** return the same drug effect estimates as **ALL**. **ML**, which does not assume that C_Q=C, performs nearly as well as **BA** and **BSR**.

Figure 5.6, the plot of $\delta_{D=Dn}$ as a function of f_M, reveals that the advantage of **BA** and **BSR** over **ML** exists only for f_M<.4. $\delta_{D=Dn}$ for **ML** converges on **BA** and **BSR** (and **ALL**) values at f_M=.4. Note that BA's performance does not depend on f_M in study 2 since C_Q=C.

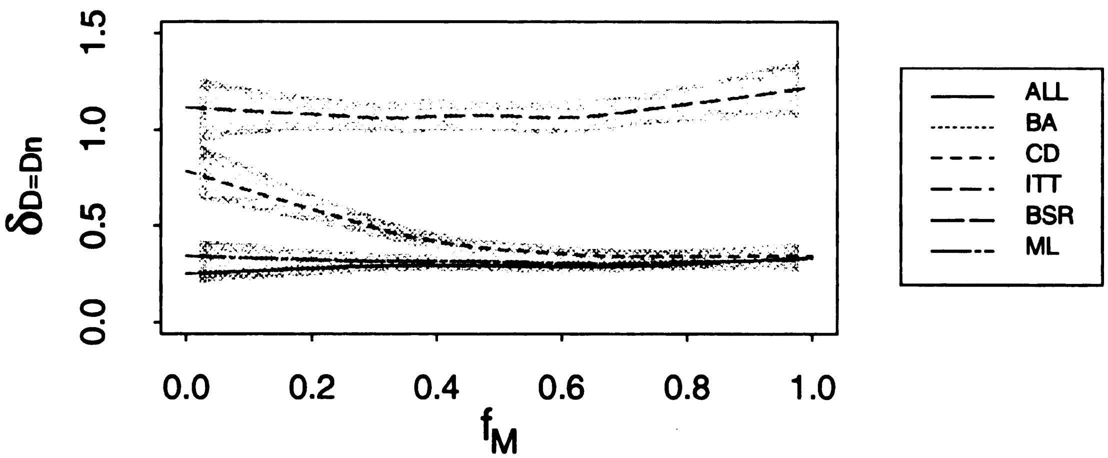


Figure 5.6. Plot of $\delta_{D=Dn}$ as a Function of f_M for Study 2. **ML** estimates of P(Y|D=D_n) converge on **BA** and **BSR** (and **ALL**) when f_M<.4.

5.4.1.2 Poor C_Q Data: Self-Reported Compliance Has No Information About True Compliance—Study 3

Figure 5.7, a boxplot of Δ_R for study 3, shows that **ML** yields the most precise estimates of drug effect of all methods when C_Q is uncorrelated with C. In contrast, the other two methods that use C_Q data are hurt by the lack of information in patient self-reported compliance. **BSR** yields the least precise estimates of P(Y|D)—worse, even, than **ITT**. (**ITT** uses a poor estimate of compliance, as it assumes perfect intake for all subjects.) **BA** yields less precise estimates of drug effect compared to study 1 (compare Figure 5.7 to Figure 5.1). Again, the precision of **CD** and **ITT** is unchanged relative to study 1. Naturally, the rank order of performance of **CD** and **ITT** is influenced by the accuracy of self-reported compliance since changes in the precision of other methods affect this ordering. Despite returning the same estimates of P(Y|D), **CD** is the second best method in studies 1 and 3, and next to worst in study 2.

A comparison of the plot of $\delta_{D=Dn}$ as a function of f_M for study 3 (Figure 5.8) to the analogous study 1 plot (Figure 5.3) offers another perspective on the influence of the self-report quality. In Figure 5.8, **BSR** overlaps with **ITT**—an extreme change compared to Figure 5.3 in which the entire 95% confidence region around **BSR** is distinct from **ITT**. **BA** has much greater error at f_M=0 when cor(C_Q,C)=0 (Figure 5.8) than when there is some information in self-reported intake (Figure 5.3). In contrast, **ML** changes little, revealing that **ML** gains as much information from the Y,C_Q data in subjects with M=0 when integrating over a flat P(C|C_Q) distribution as it does when there is perfect agreement between C_Q and C. Note that the equivalence of P(Y|D=D_n) estimates for **BSR** and **ITT** reveals that differences between Δ_R in Figure 5.7 are due to differences in P(Y|D=0) estimates.

Studies 1, 2, and 3 illustrate the importance of the quality of information in C_Q to each analysis method. As expected, **BSR**'s performance is the most sensitive to cor(C_Q,C). In study 2, where C_Q=C, **BSR** is the best method of analysis. In study 1, where $C_Q \geq C$, **BSR** is only better than **ITT**. In study 3, where $C_Q \perp C$, **BSR** performs the poorest of all methods. Since it uses some compliance data that are accurate, **BA**'s precision is less sensitive to cor(C_Q,C) than **BSR**. Although **BSR** is the best method of analysis when C_Q=C (study 2), it is only third best (or third poorest) when $C_Q \geq C$ and $C_Q \perp C$ (studies 1 and 3). In contrast, **ML**'s ability to estimate drug effect varies little. Of

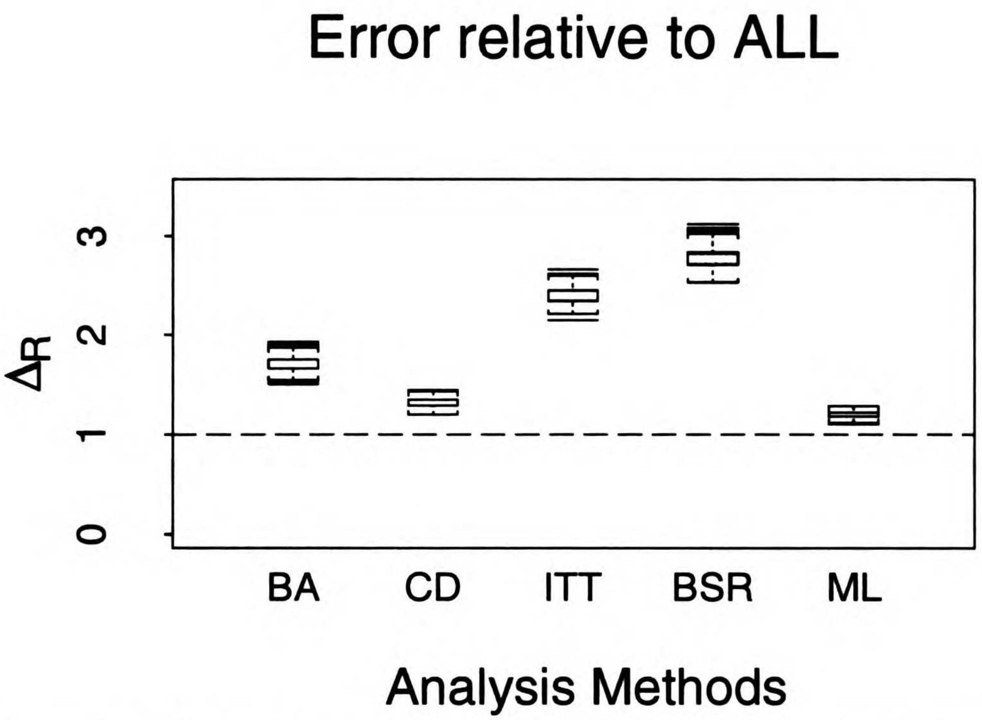


Figure 5.7. Boxplot of Δ_R for Study 3. When a patient's self reported compliance has no correlation with true intake, **ML** (which uses C_Q data) estimates drug effect better than all other methods—nearly as well as **ALL**. In contrast, **BSR** performs worse than **ITT**.

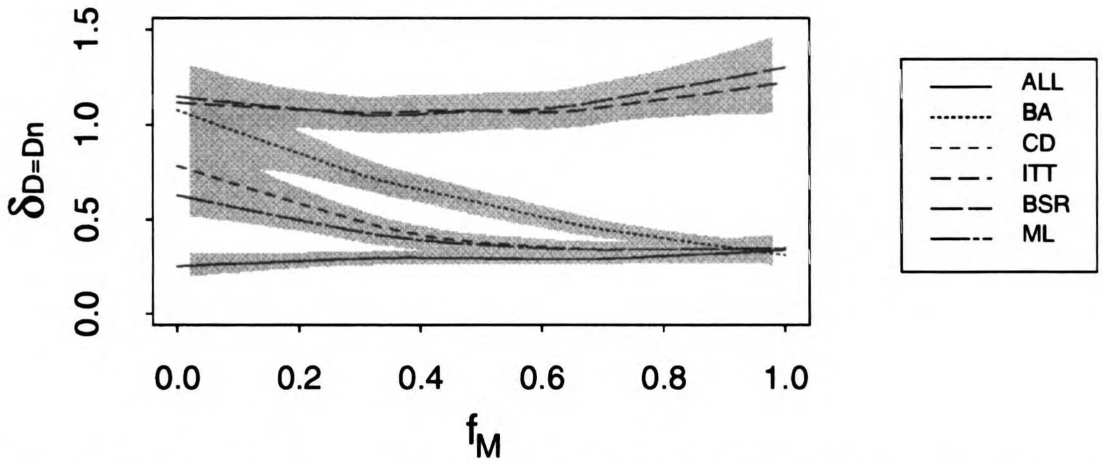


Figure 5.8. Plot of $\delta_{D=Dn}$ as a Function of f_M for Study 3. ML yields the best estimates of drug effect when cor(C_Q,C)=0, regardless of the fraction of subjects with C_M data.

course, this in itself does not imply that **ML** is a desirable method of analysis. After all, Δ_R for methods that do not use C_Q data—**CD** and **ITT**—are completely insensitive to cor(C_Q,C). However, **ML** is the best method of analysis according to studies 1, 2, and 3 because it yields the best estimates of P(Y|D) when $C_Q \geq C$ and $C_Q \perp C$ and the second best estimates when $C_Q = C$.

5.4.2 Investigation of Extremes in the Quantity of Electronically Monitored Compliance Data

This study design explores method performance under two extreme conditions of available C_M data. First, when all patients have M=1, and, second, when no patient has C_M data. For the sake of comparison, the range of all parameters, except f_M, are the same as in study 1. Refer to Table 5.6 for a summary of parameter values used in these simulations. Boldface type is used to indicate parameters drawn from different ranges than in study 1 (compare Table 5.6 to Table 5.2).

Table 5.6. Distribution of Random Parameter Values in Study 4 and Study 5.

Parameter	**Distribution (or Fixed Value)**
N	U(50,400)
f_M	**Study 4: f_M=1**
	Study 5: f_M=0
a, *b*	U(0,1), U(0,2)
ρ	U(logit(.1),logit(.5))

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 5.7.

5.4.2.1 Rich C_M Data: All Subjects Have C_M Data—Study 4

Figure 5.9, the boxplot of Δ_R for study 4, shows that three methods yield the most precise estimates of drug effect. **ML**, **CD**, and **BA** are equivalent to **ALL** because f_M=1. **ITT** and **BSR** estimates of exposure-response are unchanged compared to study 1 (compare Figure 5.9 to Figure 5.1) since neither method uses C_M data.

It is not surprising that **BA** and **CD** benefit from having complete C_M data. However, **ML** is not expected to converge on **ALL**. After all, **ML** uses all of the data—including biased C_Q data. Recall that **ML** does not estimate P(Y|D) as well as **ALL** when C_Q data are rich (see Figure 5.5).

Table 5.7. Observed Range of Parameter Values in Study 4 and Study 5.

Parameter	**Range**
N	50—400
f_M	**Study 4: f_M=1**
	Study 5: f_M=0
cor(C_Q,C)	.0593—.791
P(Y|D=0)	.101—.500
P(Y|D=D_n)	.500—.899
P(C)	P(C=0): .127—.456, P(C=.5): .228—.545, P(C=1): .207—.454

5.4.2.2 Poor C_M Data: No Subject Has C_M Data—Study 5

In the extreme case that all subjects are missing C_M data, the two methods that rely on this information—**ML** and **CD**—are expected to suffer greatly. While studies with no calibration group are irrelevant to this investigation, the results are presented to illustrate assumptions of the analysis methods that are not revealed through other designs.

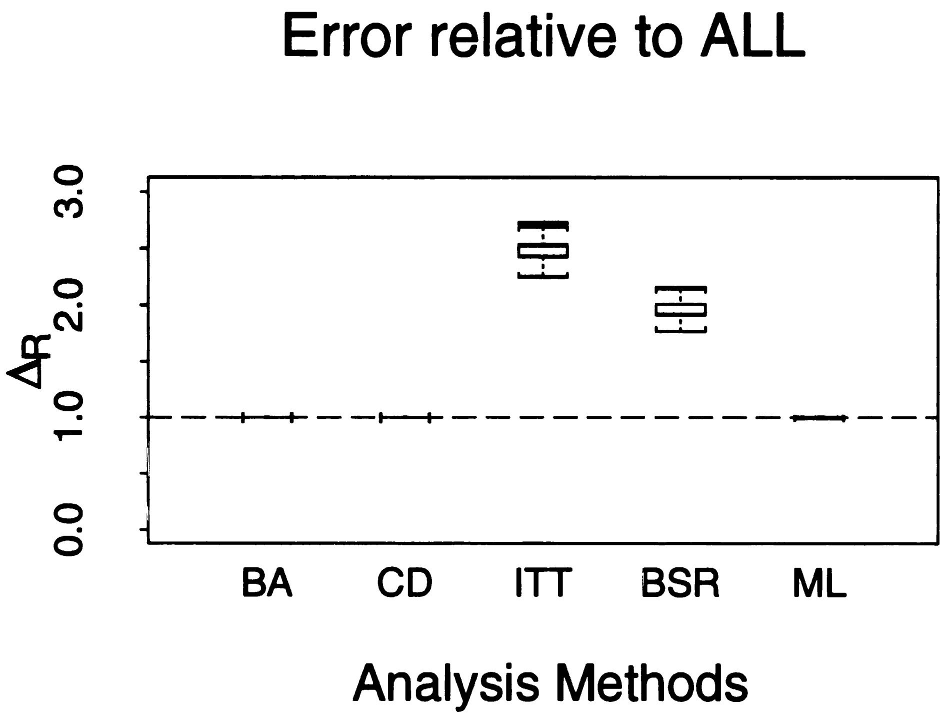


Figure 5.9. Boxplot of Δ_R for Study 4. When all subjects have C_M data, **ML** estimates P(Y|D) as well as **BA** and **CD** (and **ALL**).

Figure 5.10, a boxplot of Δ_R for study 5, shows that no method yields estimates of drug effect that approach **ALL** when f_M=0. **BA** and **BSR** yield the best estimates of

P(Y|D) and **ITT** yields the least precise estimates of drug effect. Of course, **BSR** and **ITT** estimates of drug effect are identical to study 1 (compare Figure 5.10 to Figure 5.1) since neither uses C_M data. The precision of methods that use C_M data is degraded to that of **BSR** and **ITT**—**BA** is equivalent to **BSR**, **CD** converges on **ITT**, and **ML** lies somewhere in between. The behavior of **BA** is understandable. Only **CD** and **ML** require further explanation.

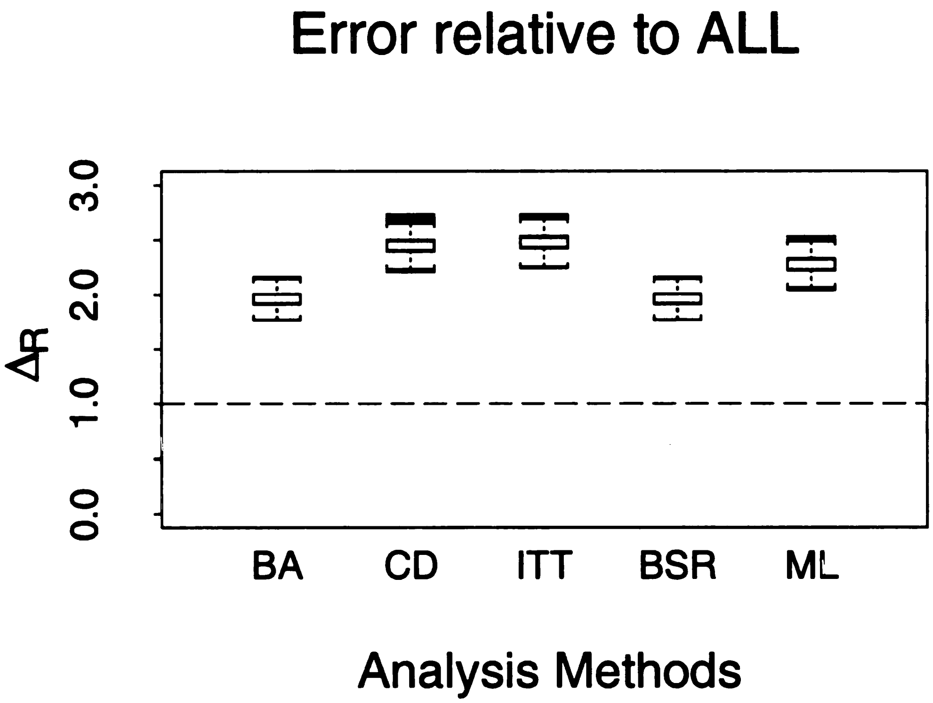


Figure 5.10. Boxplot of Δ_R for Study 5. When all C_M data are missing, no method has precision similar to **ALL**. Relatively speaking, **BA** and **BSR** yield the best estimates of drug effect. **ML** maintains its advantage over **ITT**.

To yield similar values of Δ_R, **CD**'s estimates of both P(Y|D=0) and P(Y|D=D_n) must approach **ITT**'s estimates. Their P(Y|D=0) estimates are, in fact, equivalent and their P(Y|D=D_n) estimates are similar. However, the P(Y|D=0) estimates are equivalent for a different reason than the P(Y|D=D_n) estimates are similar. First, the equivalence of P(Y|D=0) estimates is explained.

Recall that all analysis methods classify subjects with D_n=0 as having D=0 (formally, D(C,D_n=0) = 0, as described in the Appendix). Since **ITT** assumes C=1 for all subjects, it determines P(Y|D=0) using data exclusively from subjects with D_n=0. Each pseudolikelihood method (**ALL**, **BA**, **CD**, and **BSR** (and **ITT**)) yields an estimate of P(Y|D=0) identical to **ITT**'s estimate when there are no compliance data (specifically, no zero percent compliance data) apportioning subjects with D_n=1 to the D=0 group. Hence, **CD**'s estimates of P(Y|D=0) are expected to match those of **ITT** when no subject assigned to M=1 with D_n=1 has C_M=0. Under the study 5 design, *no* subject has M=1, therefore, there are no measures of C_M=0, so **CD** has the same estimates of P(Y|D=0) as **ITT**. (Since **ML** uses all of the data to estimate all parameters, it does not simply return the **ITT** estimate. Its estimates may be pulled toward it, but, ultimately, **ML** finds the best compromise between the available data, θ_{1prior} and θ_{2prior}, by integrating across the missing data.)

The explanation for the similarity between **CD** and **ITT** estimates of P(Y|D=D_n) is less straightforward. Recall that P(C) assigns each of the three values of true compliance to a subject with equal probability. Because of this, the group of subjects randomly assigned to D_n=1 should be equally comprised of subjects with C=0, C=.5, and C=1. Recall also that **ITT** estimates P(Y|D=D_n), essentially, by averaging the responses of all subjects with $D_n=1$ (since **ITT** assumes perfect compliance with the assigned regimen). By virtue of the arbitrary symmetry of $P(Y|D=0)$ and $P(Y|D=D_n)$ around .5 (refer to Figure A.6 to see that: $\text{logit}\{P(Y|D=0)\} = -\text{logit}\{P(Y|D=D_n)\}$), the mean of responses for patients assigned to $D_n=1$ is .5. That is, the mean of responses in a group of subjects equally comprised of individuals with $D=0$, $D=.5D_n$, and $D=D_n$ is .5. Of course, **ITT**'s estimate of $P(Y|D=D_n)$ may not be exactly equal to this value for all trials due to random variation in simulation.

The prior value of $P(Y|D)$ (θ_{1prior}) is equivalent to no drug effect, or formally, θ_{1prior}: $P(Y|D=0)=P(Y|D=.5D_n)=P(Y|D=D_n)$. Arbitrarily, the null drug effect was fixed to a 50 percent chance of success. When no data speak to a particular level of exposure for a given pseudolikelihood method, this prior value of $P(Y|D)$ is returned. In study 5, where $M=0$ in all N subjects, **CD** has no C_M data from subjects with $D=D_n$. The lack of $D=D_n$ data causes **CD** to simply return the prior value on $P(Y|D=D_n)$, or report $P(Y|D=D_n)=.5$. Coincidentally, this is the expected value for the **ITT** estimate. **ITT**'s estimates are not exactly the same as **CD**'s estimates because **ITT** actually analyzes simulated data while **CD** simply returns the number .5. In summary, the similarity between **CD** and **ITT**'s estimates of drug effect when $f_M=0$ is a consequence of two arbitrary choices made in study design.

Figure 5.11, a plot of $\delta_{D=Dn}$ as a function of $\text{cor}(C_Q,C)$, reveals that Figure 5.10 provides a misleading comparison of **ML** with **BSR** and **ITT**. **BSR** does not yield more precise estimates of $P(Y|D=D_n)$ than **ML** for all values of $\text{cor}(C_Q,C)$. **ML** is the most precise of all methods when $\text{cor}(C_Q,C)<.2$. **ML** does not yield more precise estimates of drug effect than **ITT** for all values of $\text{cor}(C_Q,C)$. **ML** converges on **ITT** at $\text{cor}(C_Q,C)>.6$.

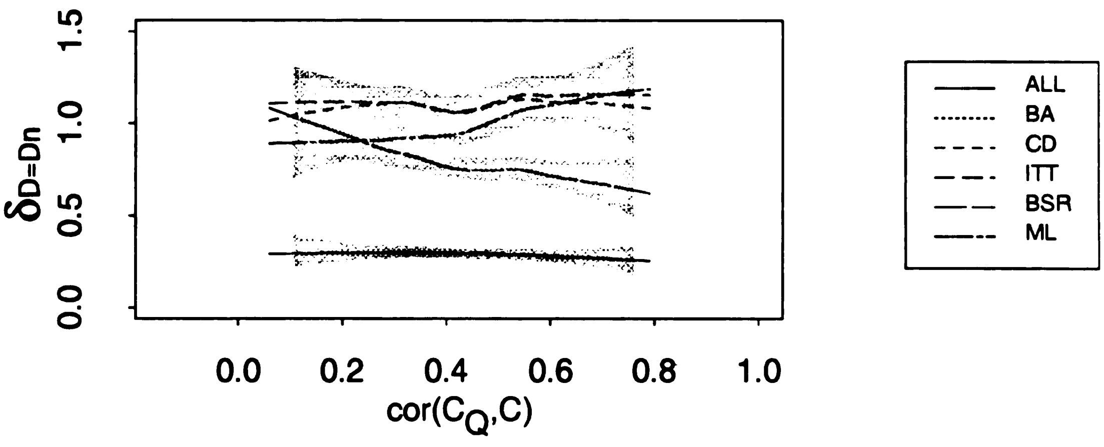


Figure 5.11. Plot of $\delta_{D=Dn}$ as a Function of cor(C_Q,C) for Study 5. This plot illustrates the influence of the prior P(C|C_Q) on **ML** performance in a situation where it is expected to noticeably impact the method—there are no C_M data. Here, θ_{2prior} assumes zero correlation between C_Q and C. When there is, in fact, no information on C in C_Q (cor(C_Q,C)<.2), **ML** yields the best estimates of P(Y|D=D_n). However, as cor(C_Q,C) increases (has less agreement with the prior) **ML**'s performance approaches that of **ITT**.

ML's error increases dramatically as the accuracy of C_Q improves—a counterintuitive result. One might expect **ML** to perform better as cor(C_Q,C) increases. But because **ML** has no calibration data, it has no anchor for estimating P(C|C_Q). Consequently, **ML** relies heavily on the prior P(C|C_Q) in this case. Recall that θ_{2prior} states that there is no relationship between C_Q and C. When the correlation between C_Q and C is, in fact, near zero, the **ML** approach is most appropriate since the underlying assumption of the prior—no correlation between C_Q and C—is borne out by the data. However, for high cor(C_Q,C), the prior advises **ML** to ignore the (valid) information in the C_Q data and **ML**'s precision becomes equivalent to **ITT**.

5.5 Determination of the Influence of Trial Design on Method Performance—Non-Placebo Controlled Studies

This study is carried out to determine if simulating placebo-controlled trials confers an unfair advantage on **ML** relative to other analysis methods. To understand why **ML** is suspect, one must remember that estimation of P(Y|D=0) is accurate for all analysis methods under the placebo controlled design because at least half of all subjects contribute to the estimate—half are assigned to D_n=0 and D(C,D_n=0)=0. Of all analysis methods, however, **ML** is the only one that estimates all parameters simultaneously. Perhaps **ML**'s ability to estimate one parameter very well serves as a helpful constraint in estimation?

To explore this question, studies 1-5 are carried out with only the unit dose group (D_n: (1))—none of the N subjects in each trial are assigned placebo. These studies are named with the superscript "$^{Dn=1}$" indicating that only a unit dose is assigned.

5.5.1 Investigation of Performance Over a Wide Range of Parameter Space—Study $1^{Dn=1}$

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 5.8.

Figure 5.12, the boxplot of Δ_R in study $1^{Dn=1}$, shows results analogous to study 1 (compare to Figure 5.1). **ML** has the greatest estimation precision and the rank order of method performance remains unchanged. However, it also shows that **ITT**, **BSR**, and **BA** have much less precision (greater error) than in study 1. The reason for this change in performance is apparent in Figure 5.13—the plots of $\delta_{D=0}$ and $\delta_{D=Dn}$ estimates versus cor(C_Q,C) for this study. In study 1, all methods yield estimates of P(Y|D=0) with precision equivalent to **ALL**, hence this result is not presented. Figure 5.13 shows that in the absence of a placebo group, **ITT**, **BSR**, and **BA** yield low precision P(Y|D=0) estimates with error equivalent in magnitude to that of P(Y|D=D_n) estimates. The placebo group has a particularly beneficial effect on **ITT**, **BSR**, and **BA** estimates for the following reason.

Table 5.8. Observed Range of Parameter Values in Study $1^{Dn=1}$.

Parameter	**Range**
N	50—400
f_M	.00273—.997
cor(C_Q,C)	.0546—.801
P(Y|D=0)	.100—.500
P(Y|D=D_n)	.500—.900
P(C)	P(C=0): .2-.5, P(C=.5): .188-.441, P(C=1): .154-.471

In study 1, at least 50% of the data contribute to P(Y|D=0) estimation—data from subjects with D_n=0. In study $1^{Dn=1}$ there is no such advantage for P(Y|D=0) estimation. **ITT** returns the prior value on P(Y|D=0) as an estimate of drug effect since **ITT** determines P(Y|D=0) from D_n=0 subjects only and there is no placebo group. **BSR** estimates P(Y|D=0) using data from subjects with D_n=1 and C_Q=0 only. The lack of D_n=0 data reduces the precision of this estimate. **BA**'s estimates of P(Y|D=0) are hurt by the same factors acting on **BSR**'s estimates. However, since **BA** has some C_M data, its precision is less sensitive to this study design.

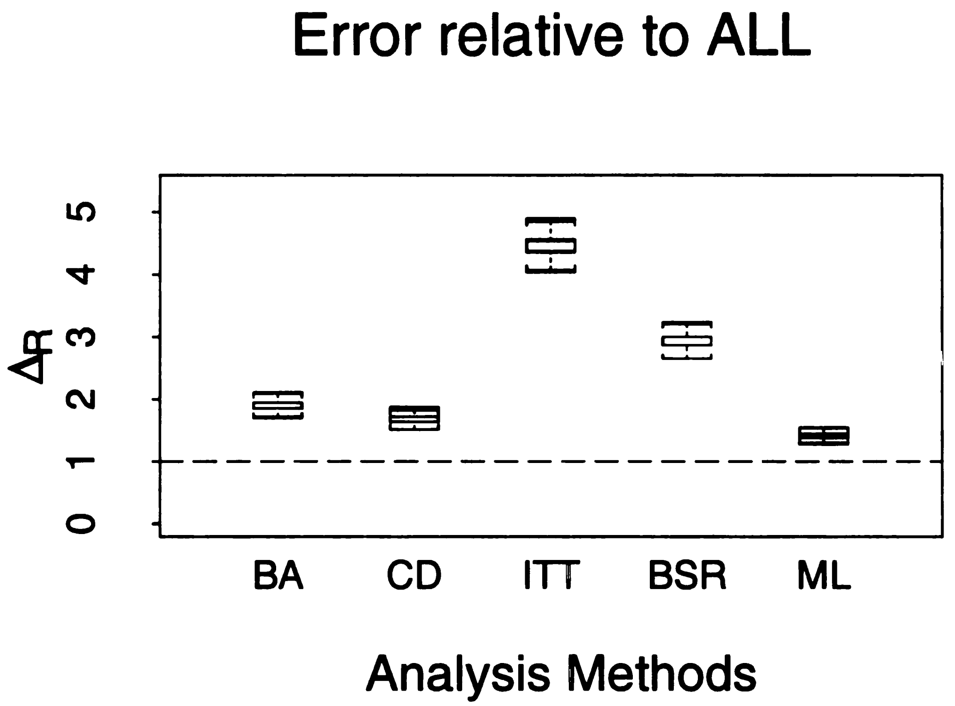


Figure 5.12. Boxplot of Δ_R for Study $1^{Dn=1}$. The methods exhibit the same relative performance under a non-placebo-controlled trial design as for placebo-controlled trials (compare to **Figure 5.1**). The magnitude of error in estimates of P(Y|D) is much higher, however, for **ITT** and **BSR** when trials lack a placebo group.

There is a slight difference between **BSR**'s estimates of P(Y|D=0) and P(Y|D=D_n) in study $1^{Dn=1}$ that warrants mention. At cor(C_Q,C)>.7, **BSR**'s P(Y|D=0) estimates are nearly equivalent to **ALL** but **BSR** yields poorer estimates of P(Y|D=D_n) than **ALL**. This

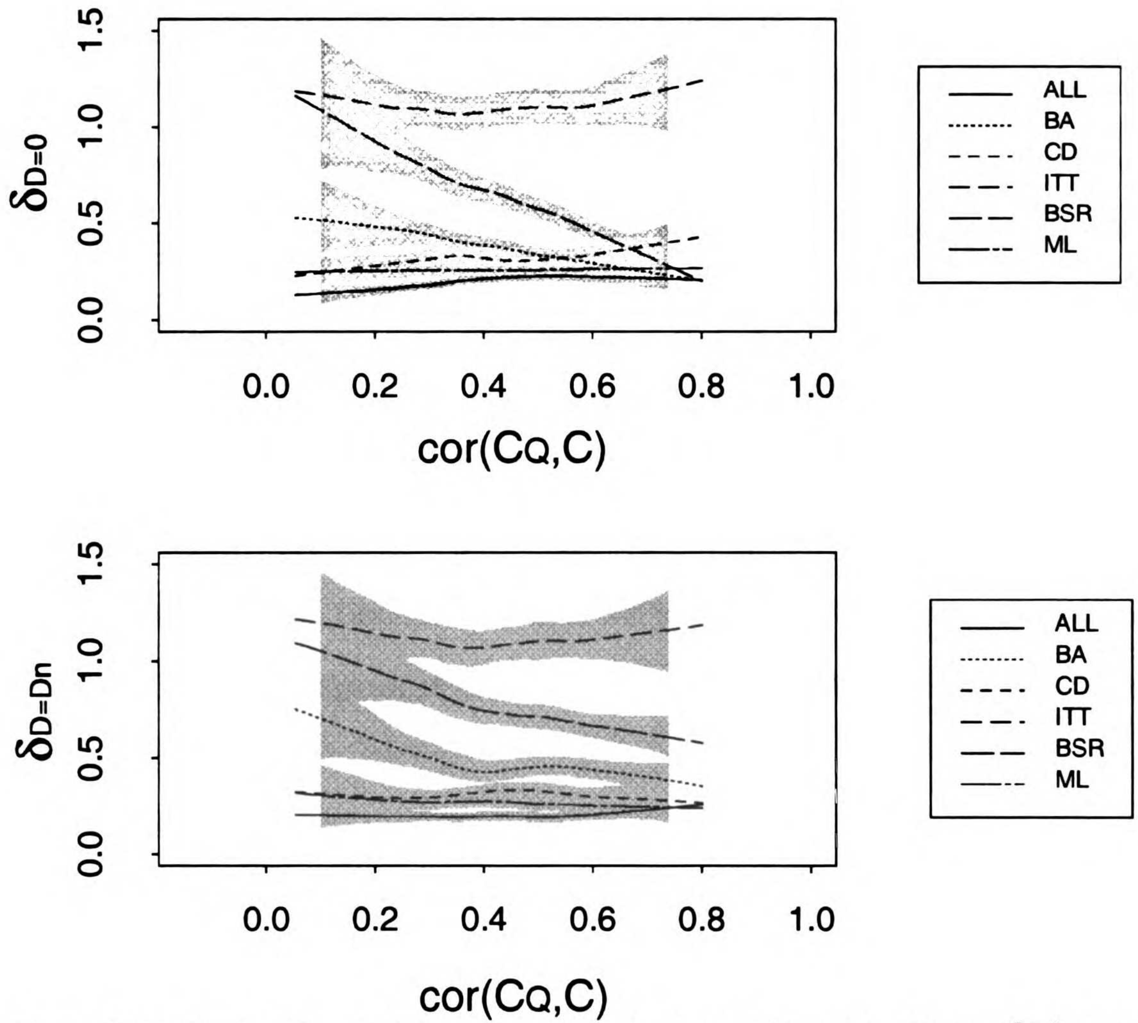


Figure 5.13. Plots of $\delta_{D=0}$ and $\delta_{D=Dn}$ as a Function of cor(C_Q,C) for Study $1^{Dn=1}$. Error in estimates of P(Y|D=0) and P(Y|D=D_n) is similar for each method when there is no placebo group assuring that at least half of the data contribute to P(Y|D=0) estimation. Note that the difference in **BSR**'s ability to estimate P(Y|D=0) and P(Y|D=D_n) at cor(C_Q,C)>.7 reflects the difference in accuracy of a self-reported perfect compliance versus a self-reported zero compliance.

simply reflects a discrepancy in the accuracy of C_Q=0 versus the accuracy of C_Q=1 as a measure of C.

Recall that C_Q is simulated from a model, $P(C_Q|C)$, that is constructed by weighting three distributions: $C_{Q,Q=C}$, $C_{Q,Q\geq C}$, and $C_{Q,Q\perp C}$ (see Table A.1). The $C_{Q,Q\geq C}$ model states that a subject who self reports zero compliance, in fact, has zero compliance, while a subject who self-reports one hundred percent compliance has some probability of having zero, fifty, and one hundred percent compliance. Thus, in $C_{Q,Q\geq C}$, $C_Q=0$ is a more valid measure of compliance than $C_Q=1$(Caron 1985). $C_{Q,Q=C}$ and $C_{Q,Q\perp C}$ are equally valid for all values of C_Q, as C_Q is either entirely accurate or entirely uncorrelated with C in these models. Consequently, as long as $WT_{Q\geq C}$ is greater than 0, this discrepancy in $P(Y|D=0)$ and $P(Y|D=D_n)$ estimation by **BSR** will be observed. Note that the correlation between C_Q and C is computed on data simulated for each trial, thus it is only an indirect indicator of $P(C_Q|C)$.

5.5.2 Investigations of Performance at Specific Locations in Parameter Space

Studies 2-5 evaluate method performance given extremes in the quality and quantity of compliance data under a placebo-controlled design. One can imagine that there may be an interaction between the quality and quantity of compliance data and study design. For example, favorable properties conferred on methods by virtue of estimating $P(Y|D=0)$ well (because placebo is assigned to half of all subjects) may mask unfavorable properties due to having little compliance data. To determine the magnitude of this effect, studies 2-5 are repeated with all subjects assigned to the unit dose only. Comparison of these results to the results of the analogous placebo-controlled study elucidates the interaction between study design and compliance information.

5.5.2.1 Investigation of Extremes in the Quality of Self-Reported Compliance Data

This study design explores the performance of methods given two extremes in self-reported compliance accuracy. First, the performance when all patients report compliance accurately, or $\mathrm{cor}(C_Q,C)=1$, is considered (rich C_Q data). The second simulation study addresses the situation in which there is little correlation between a patient's self-reported compliance and his true intake (poor C_Q data). For the sake of comparison, in these studies, N, f_M, and ρ range as in study 2 and study 3. The only difference is that half of the subjects in study 2 and study 3 are assigned drug ($D_n=1$) whereas in study $2^{Dn=1}$ and study $3^{Dn=1}$, all N subjects are assigned drug.

5.5.2.1.1 Rich C_Q Data: Self-Reported Compliance is Accurate—Study $2^{Dn=1}$

The purpose of this study is to investigate the interaction between the quality of self-reported compliance data and study design. This study is identical to study 2 except that it is not placebo-controlled. Therefore, the only way for subjects to have zero exposure to drug is to have zero compliance.

Observed values of certain parameters in the data actually simulated in study $2^{Dn=1}$ are given in Table 5.9.

Figure 5.14, the boxplot of Δ_R in study $2^{Dn=1}$, shows results consistent with the analogous placebo-controlled study (study 2)—**BA** and **BSR** have estimation error equivalent to **ALL** and the rank order of method performance is maintained (compare to Figure 5.5). The most striking discrepancy is the decrease in **ITT**'s estimation precision. This decrease is of no concern, however, because it is the worst performing method of analysis in both study 2 and study $2^{Dn=1}$.

Table 5.9. Observed Range of Parameter Values in Study $2^{Dn=1}$.

Parameter	Range
N	51—400
f_M	0—1
cor(C_Q,C)	1
P(Y|D=0)	.100—.499
P(Y|D=D_n)	.501—.900
P(C)	P(C=0): .173-.472, P(C=.5): .208-.558, P(C=1): .182-.435

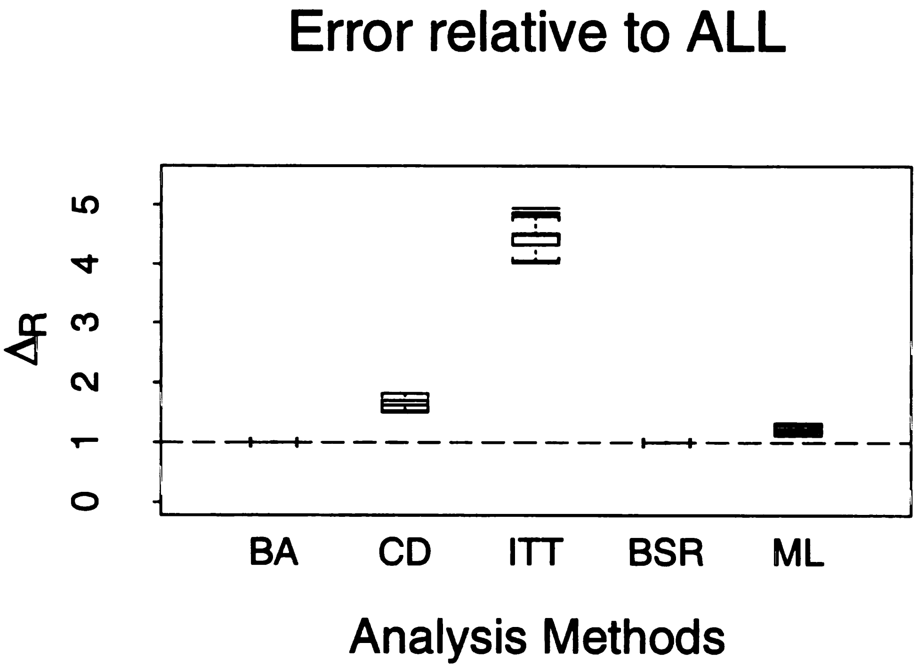


Figure 5.14. Boxplot of Δ_R for Study $2^{Dn=1}$. Comparison with results of the analogous placebo-controlled study (Figure 5.5) reveals that **ML**'s performance is unaffected by the single dose design when self-reported compliance is accurate.

5.5.2.1.2 Poor C_Q Data: Self-Reported Compliance Has No Information About True Compliance—Study $3^{Dn=1}$

This study is identical to study 3, except that it is not placebo-controlled. Observed values of certain parameters in the data actually simulated in study $3^{Dn=1}$ are given in Table 5.10.

Figure 5.15, the plot of Δ_R in study $3^{Dn=1}$, shows the same rank order of performance as in the analogous placebo-controlled study (compare to Figure 5.7). However, **ITT**'s estimation precision is disproportionately poorer—it more closely approximates **BSR**'s estimation precision. **ITT** is more sensitive to the dosing design than **BSR** because the design reduces the amount of data it has to estimate P(Y|D=0). **ML** also performs poorer relative to study 3. This reflects the slight benefit gained by being able to estimate P(Y|D=0) well in a simultaneous estimation of all parameters.

Table 5.10. Observed Range of Parameter Values in Study $3^{Dn=1}$.

Parameter	Range
N	51—400
f_M	0—1
cor(C_Q,C)	-.165—.307
P(Y|D=0)	.100—.499
P(Y|D=D_n)	.501—.900
P(C)	P(C=0): .173-.472, P(C=.5): .208-.558, P(C=1): .182-.435

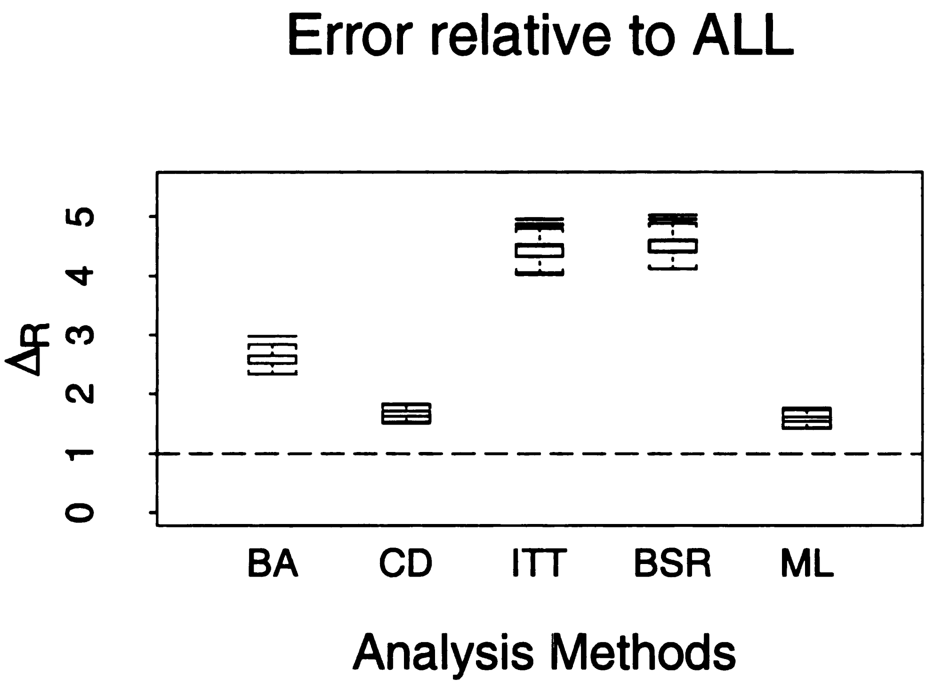


Figure 5.15. Boxplot of Δ_R for Study $3^{Dn=1}$. Comparison with results of the analogous placebo-controlled study (Figure 5.7) reveals that **ML**'s performance is slightly affected by the single dose design when C_Q is uncorrelated with C. However, it is still the best method of analysis. **ITT**'s performance changes drastically—it is nearly as bad as **BSR**—indicating that this method is most sensitive to the dosing design.

5.5.2.2 Investigation of Extremes in the Quantity of Electronically Monitored Compliance Data

5.5.2.2.1 Rich C_M Data: All Subjects Have C_M Data—Study $4^{Dn=1}$

As one may expect from the results of the previous single dose studies, **BA**, **CD**, and **ML** yield the most precise estimates of Δ_R and **ITT** and **BSR** suffer large estimation errors under this design. These results are trivial, and are not further elaborated on.

5.5.2.2.2 Poor C_M Data: No Subject Has C_M Data—Study $5^{Dn=1}$

This study is identical to study 5 except that it is not placebo controlled.

Observed values of certain parameters in the data actually simulated in study $5^{Dn=1}$ are given in Table 5.11.

Table 5.11. Observed Range of Parameter Values in Study $5^{Dn=1}$.

Parameter	Range
N	50—400
f_M	0
cor(C_Q,C)	.0593—.791
P(Y|D=0)	.101—.500
P(Y|D=D_n)	.500—.899
P(C)	P(C=0): .127-.456, P(C=.5): .228-.545, P(C=1): .207-.454

A comparison of Figure 5.16, the boxplot of Δ_R in study $5^{Dn=1}$, to Figure 5.10, the boxplot of Δ_R in the analogous placebo-controlled design, reveals that the placebo-controlled design does not influence the rank order of method performance.

5.5.3 Summary

In summary, the relative performance of methods is equivalent for placebo-controlled and single dose designs. Therefore, there is no significant interaction between study design and the quality and quantity of compliance information for the range of parameters explored.

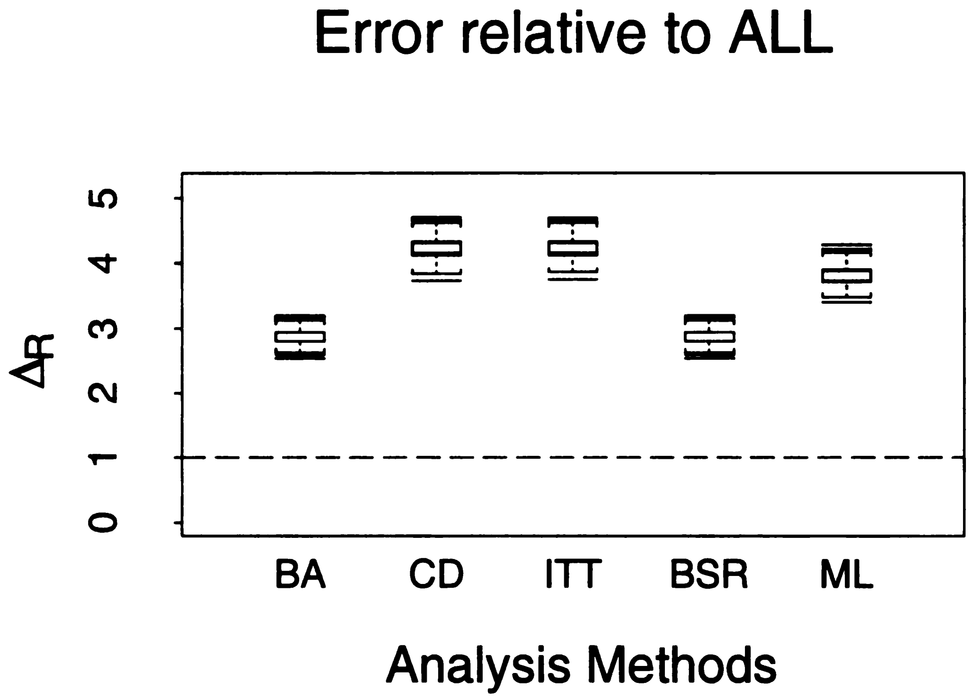


Figure 5.16. Boxplot of Δ_R for Study $5^{Dn=1}$. Comparison with results of the analogous placebo-controlled study (Figure 5.10) reveals that **ML**'s performance is unaffected by the single dose design when there is no C_M data.

5.6 Determination of the Influence of the Prior $P(C|C_Q)$ on Method Performance

A weak and uninformative prior value of $P(C|C_Q)$ is assumed in studies 1-5. It is of interest to determine the relationship, if any, between method performance and this prior information.

To investigate this question, the prior $P(C|C_Q)$ is fixed to perfect correlation between C and C_Q—an extreme change from the prior of $C_Q \perp C$ used previously. For the sake of comparison, the study designs are identical to those already discussed, except, of course, for the value of θ_{2prior}. The superscript "$^{\theta 2prior:CQ=C}$" is appended to the study name to indicate that this is the only altered study parameter.

The results of two studies are reported. Study $1^{\theta 2prior:CQ=C}$ is reported because it demonstrates θ_{2prior}'s influence over general parameter space. Study $5^{\theta 2prior:CQ=C}$ is reported because study 5 raises concern about θ_{2prior}. The rest of the results show no appreciable influence of θ_{2prior} on method performance and are not discussed.

Table 5.12 lists the parameter values common to these two studies. Note that boldface type is used to indicate the fixed values differing between all previous studies and the "$^{\theta 2prior:CQ=C}$" studies (compare Table 5.12 to Table 5.1).

Table 5.12. List of Fixed Parameter Values Common to All Studies Investigating Sensitivity to the Prior P(C|C$_Q$) (Study $1^{\theta 2prior:CQ=C}$ and Study $5^{\theta 2prior:CQ=C}$).

Parameter	**Fixed Value**
N_{lhs}	100 ×5
P(C)	P(C=0)=P(C=.5)=P(C=1)=1/3
λ_1	.5/3
λ_2	.5/6
θ_{1prior}	$P(Y|D=0)=P(Y|D=.5D_n)=P(Y|D=D_n)= .5$
$\boldsymbol{\theta_{2prior}}$	**$P(C=k|C_Q=j)=1$ for $j=k$, $P(C=k|C_Q=j)=0$ for $j\neq k$**
$P(C_M|C)$	$C_{M,M=C}$ (refer to Figure A.5)

5.6.1 Investigation of Performance Over a Wide Range of Parameter Space— Study $1^{\theta 2\mathrm{prior}:CQ=C}$

Refer to Table 5.13 for a summary of parameter values used in these simulations. Observed values of certain parameters in the data actually simulated are given in Table 5.14.

Table 5.13. Distribution of Random Parameter Values in Study $1^{\theta 2\mathrm{prior}:CQ=C}$.

Parameter	Range
N	U(50,400)
f_M	U(0,1)
a, *b*	U(0,1), U(0,2)
ρ	U(logit(.1),logit(.5))

Table 5.14. Observed Range of Parameter Values in Study $1^{\theta 2\mathrm{prior}:CQ=C}$.

Parameter	Range
N	50—400
f_M	.00273—.997
cor(C_Q,C)	.0546—.801
P(Y|D=0)	.100—.500
P(Y|D=D_n)	.500—.900
P(C)	P(C=0): .2—.5, P(C=.5): .1875—.441, P(C=1): .154—.471

The boxplot of Δ_R, and the plots of $\delta_{D=Dn}$ as a function of cor(C_Q,C), f_M, and ρ for study $1^{\theta2prior:CQ=C}$ look similar to the corresponding plots made for study 1. (Compare Figure 5.17 to Figure 5.1 to see the Δ_R result. The rest are not shown because the result is trivial.). This suggests that the prior value of $P(C|C_Q)$ has little effect on ML's performance across the wide range of parameter values explored.

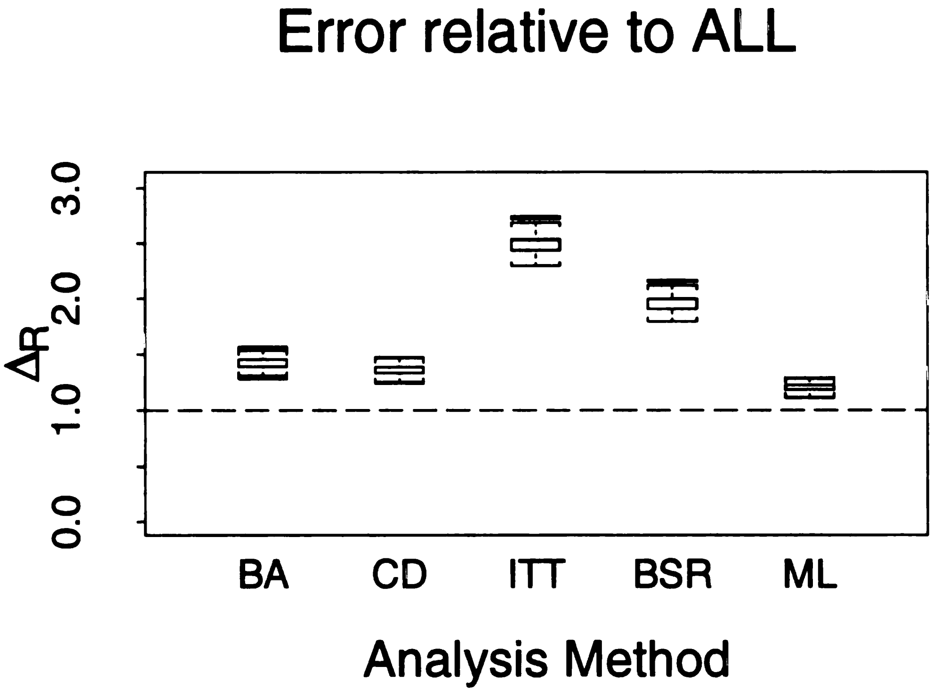


Figure 5.17. Boxplot of Δ_R for Study $1^{\theta2prior:CQ=C}$. ML's performance is unaffected by a drastic change in the prior $P(C|C_Q)$. Δ_R for this study—in which θ_{2prior} is of perfect correlation between C and C_Q—is similar to Δ_R of a study in which θ_{2prior} is of no correlation between C and C_Q (compare Figure 5.17 to Figure 5.1.)

5.6.2 Investigation of Performance at A Specific Location in Parameter Space

5.6.2.1 No Subject Has C_M data—Study $5^{\theta 2prior:CQ=C}$

The interaction between the prior $P(C|C_Q)$ and quantity of C_M data is now explored. Refer to Table 5.15 for a summary of parameter values used in these simulations.

Table 5.15. Distribution of Random Parameter Values in Study $5^{\theta 2prior:CQ=C}$.

Parameter	**Distribution (or Fixed Value)**
N	U(50,400)
f_M	0
a, *b*	U(0,1), U(0,2)
ρ	U(logit(.1),logit(.5))

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 5.16.

A comparison of Figure 5.18, the boxplot of Δ_R for study $5^{\theta 2prior:CQ=C}$, to Figure 5.10, the boxplot of Δ_R for study 5, illustrates that **ML** yields better drug effect estimates when θ_{2prior} assumes perfect correlation between C_Q and C than when it assumes no correlation between C_Q and C. Figure 5.19, the plot of $\delta_{D=Dn}$ as a function of cor(C_Q,C) for study $5^{\theta 2prior:CQ=C}$, shows that **ML**'s drug effect estimates improve as cor(C_Q,C) increases. This supports a finding of study 5 (compare to Figure 5.11). Although the

Table 5.16. Observed Range of Parameter Values in Study $5^{\theta 2prior:CQ=C}$.

Parameter	**Range**
N	50—400
f_M	0
cor(C_Q,C)	.0593—.791
P(Y|D=0)	.101—.500
P(Y|D=D_n)	.500—.899
P(C)	P(C=0): .127—.456, P(C=.5): .228—.545, P(C=1): .207—.454

results appear very different, they both reflect θ_{2prior}'s influence on **ML** when there is no C_M data. Here, **ML**'s drug effect estimates improve as cor(C_Q,C) increases, in contrast to Figure 5.11, where they become poorer as cor(C_Q,C) increases.

5.6.3 Summary

In summary, the prior on P(C|C_Q) only influences **ML** when there are no C_M data. Study $5^{\theta 2prior:CQ=C}$ suggests that a prior P(C|C_Q) of no correlation between C and C_Q is a good choice for these simulation studies. Since it penalizes **ML** more than θ_{2prior} of perfect agreement between C and C_Q, it is a conservative direction in which to err. Futhermore, since it gives rise to an unusual result, it signals to the data analyst that something is amiss.

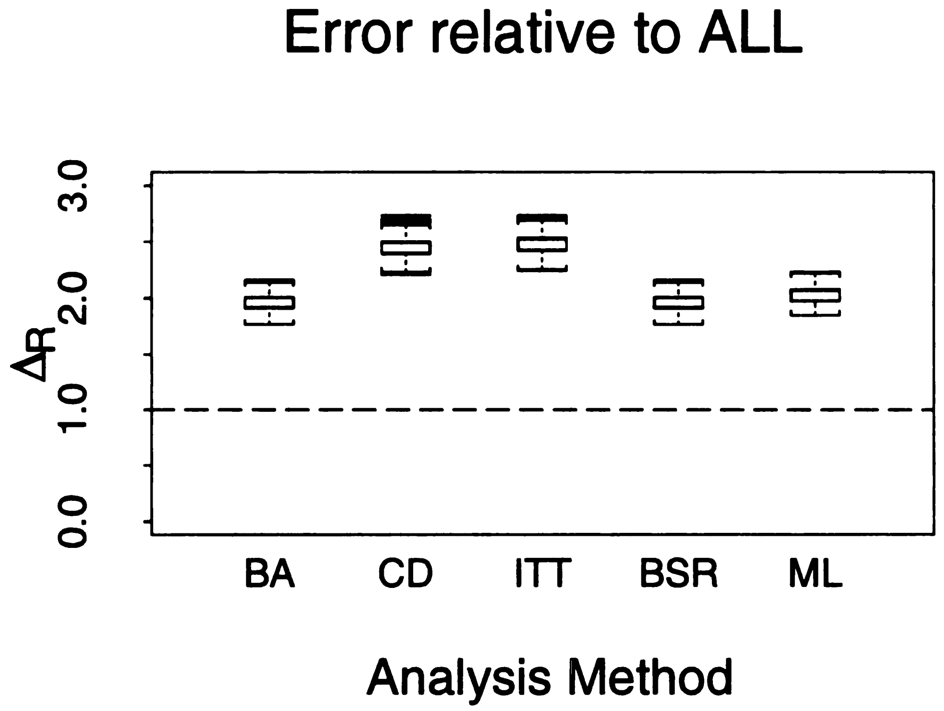


Figure 5.18. Boxplot of Δ_R for Study $5^{\theta 2prior:CQ=C}$. ML's drug effect estimates are more precise when θ_{2prior} assumes perfect agreement between C_Q and C relative to the analogous study in which θ_{2prior} assumes no correlation between C_Q and C. (Compare Figure 5.18 to Figure 5.10.) Thus, the prior $P(C|C_Q)$ used is conservative in that it penalizes **ML** relative to other methods.

5.7 Determination of the Influence of P(C) on Method Performance

The assumption of the investigations presented is that patient exposure to drug is an experimental treatment, assigned via stratified sampling through P(C). It is known that the precision of estimators computed on stratified samples is influenced by the number of experimental units apportioned to each strata. Therefore, it is of interest to determine the extent to which the choice of P(C) influences method precision. To investigate this question, simulation studies are carried out varying the distribution of P(C) between trials.

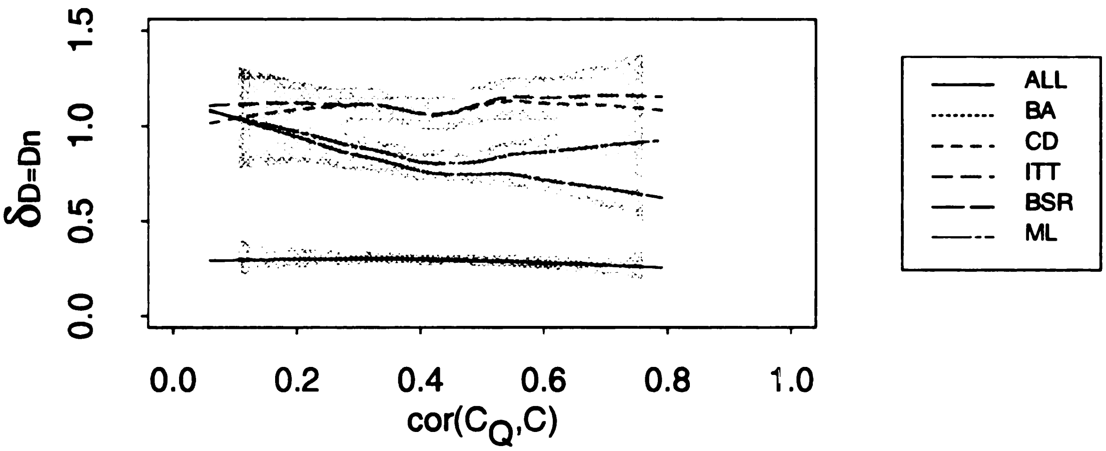


Figure 5.19. Plot of $\delta_{D=Dn}$ as a Function of cor(C_Q,C) for Study $5^{\theta 2prior:CQ=C}$. This plot illustrates the influence of the prior P(C|C_Q) on **ML** in a situation where it noticeably impacts the method—when there are no C_M data. Here, θ_{2prior} assumes perfect correlation between C_Q and C. **ML**'s estimates of P(Y|D=D_n) improve as cor(C_Q,C) better agrees with the prior P(C|C_Q) (as cor(C_Q,C) increases). Compare to Figure 5.11 which shows **ML**'s performance worsening as cor(C_Q,C) increases when the prior P(C|C_Q) assumes no correlation between C and C_Q.

In studies 1-5, P(C) assigns each patient to one of the three categories of true compliance with equal probability. Here, two extreme models for P(C) are considered—one in which compliance is skewed toward nominal intake and one in which compliance is skewed toward zero intake. These distributions are denoted by "$C_{0<.5<1}$" and "$C_{0>.5>1}$", respectively. The exact probabilities of assigning C=0, C=.5, and C=1 by $C_{0<.5<1}$ and $C_{0>.5>1}$ are given in Table 5.17. Note that the distribution of P(C) used in studies 1-5 is indicated in this chart for comparative purposes and denoted $C_{0=.5=1}$.

Table 5.17. Probability that Patients are Assigned C=0, C=.5, and C=1 for the Distributions $C_{0=.5=1}$, $C_{0<.5<1}$, and $C_{0>.5>1}$.

	P(C=0)	P(C=.5)	P(C=1)
$C_{0=.5=1}$	1/3	1/3	1/3
$C_{0<.5<1}$	1/10	1/5	7/10
$C_{0>.5>1}$	7/10	1/5	1/10

Attention is focused on the influence of P(C) by fixing all parameter values in each study rather than performing Latin Hypercube Sampling. Because of this, the number of replications performed is smaller than in previous studies—one hundred data sets are simulated given each chosen set of parameters.

It is of interest to determine if there is any interaction between P(C) and f_M at an extreme values of f_M since **ML** is the method under scrutiny and study 5 shows that its performance is most strongly affected by a lack of C_M data. A poor tool for measuring C_Q is used to highlight the influence of P(C) and f_M on **BA**, as well. (A good self-report tool would mask this interaction.) **BSR** performance will not be judged in these studies because it is unfair to evaluate **BSR** when there is so little information in C_Q. It is only included to serve as benchmark for comparison.

In this investigation, two studies are carried out—$C_{0<.5<1}$ and $C_{0>.5>1}$ are each run with f_M=.2. Table 5.18 lists the values of fixed parameters common to these simulations. Table 5.19 indicates which P(C) distribution is used in each study. The studies are not placebo controlled.

Table 5.18. Fixed Parameter Values Common to Variable P(C) Studies.

Parameter	**Value**
λ_1	.5/3
λ_2	.5/6
θ_{1prior}	$P(Y\|D=0)=P(Y\|D=.5D_n)=P(Y\|D=D_n)= .5$
θ_{2prior}	$P(C=k\|C_Q=j)=1/3$ for j,k {0,.5,1}
N	100
$P(C_Q\|C)$	$C_{Q,Q\perp C}$ (See Table A.1)
$P(C_M\|C)$	$C_{M,M=C}$ (See Figure A.5)
ρ	logit(.2)

Table 5.19. Fixed Parameter Values Varying Between Variable P(C) Studies.

Study A		**Study B**	
Parameter	**Value**	**Parameter**	**Value**
P(C)	$C_{0<.5<1}$	P(C)	$C_{0>5>1}$

Observed values of certain parameters in the data actually simulated are given in Table 5.20. Note that although $P(C_Q|C)$ is fixed to $C_{Q,Q\perp C}$, by luck of the draw, $cor(C_Q,C)$ is not always equal to 0. Likewise, the distribution of observed C varies around the simulation probabilities in $C_{0<.5<1}$ and $C_{0>5>1}$.

Table 5.20. Parameter Values Observed in Simulated Data for Variable P(C) Studies.

	Study A	**Study B**
Parameter	**Range (or Value)**	**Range (or Value)**
$cor(C_Q,C)$	-.290—.240	-.203—.219
P(C=0)	.04—.18	.56—.82
P(C=.5)	.12—.34	.11—.30
P(C=1)	.57—.80	.03—.17
N	100	100
f_M	.2	.2
$cor(C_M,C)$	1	1
P(Y|D=0)	.200	.200
$P(Y\|D=D_n)$	.800	.800

The results of studies A and B are presented in Figures 5.20 and 5.21. The study error for each method over the 100 simulations performed is plotted.

Figure 5.20 shows that **ML** yields the best drug effect estimates when compliance is good (here, 57 to 80 percent of subjects are perfect compliers). **ML** is more precise than **CD** because it uses response data from all subjects. Although the assumption of

ITT—all patients comply with their assigned regimen—is more true than in studies in which P(C) is uniform (studies 1-5), ITT still yields poor estimates of drug effect.

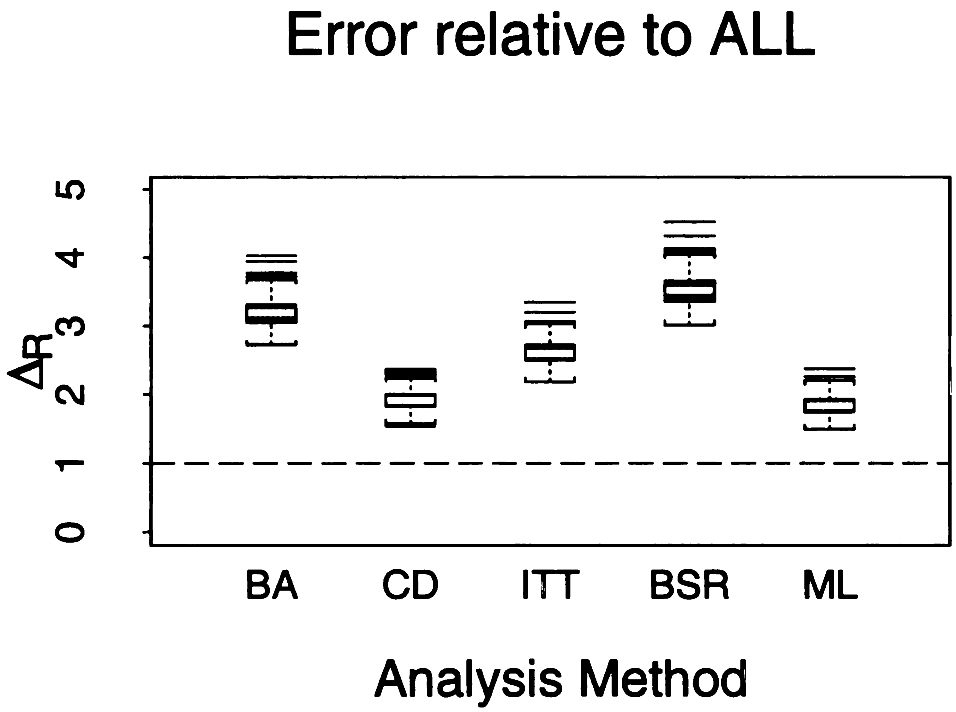


Figure 5.20. Boxplot of Δ_R for Study A. When P(C) is skewed toward good compliance, **ML** yields the best estimates of P(Y|D). Although **ITT**'s assumption of perfect compliance is closer to the truth in this study than in studies with uniform P(C), **ITT** still yields noticeably biased estimates of exposure-response.

Figure 5.21 shows that when compliance is poor (here, 56 to 82 percent of subjects take none of the prescribed drug), all methods but ITT perform the same as in studies with P(C) skewed toward good intake. **ITT** suffers because it assumes perfect compliance—an assumption that is not robust to variability in P(C).

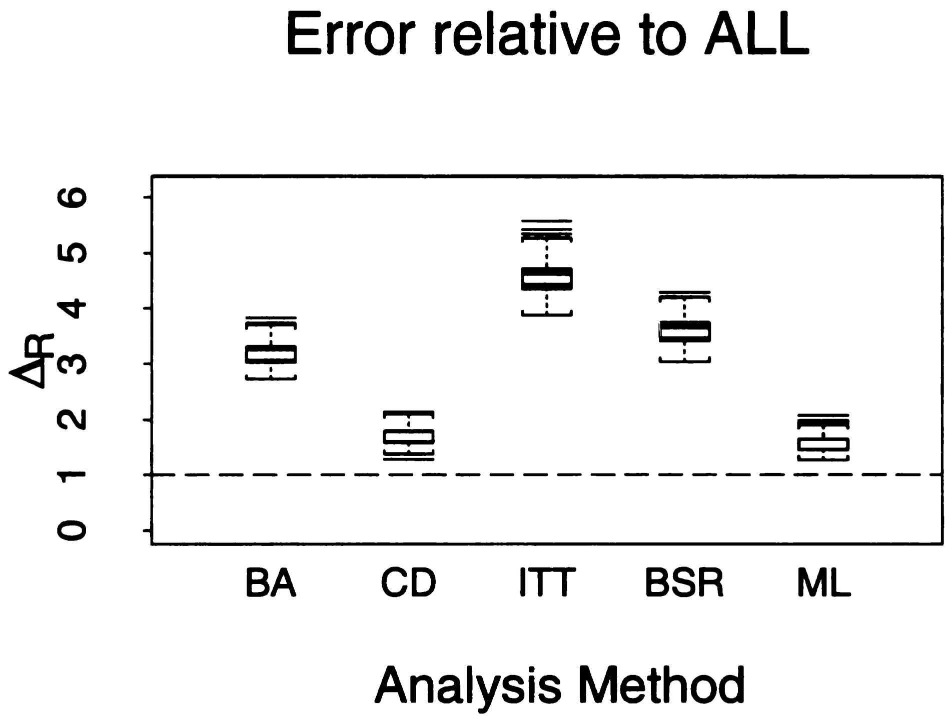


Figure 5.21. Boxplot of Δ_R for Study B. When P(C) is skewed toward poor compliance, all methods but **ITT** perform the same as in studies with P(C) skewed toward good intake. This highlights the danger of **ITT**'s assumption of perfect compliance.

In summary, since relative method performance is unchanged in these studies, the distribution of P(C) arbitrarily selected for studies 1-5 is a reasonable choice. **ML** gains no advantage over other methods of analysis by virtue of using a uniform P(C). If anything, uniform P(C) offers a conservative estimate of **BA** performance.

These studies highlight that, short of having all subjects comply perfectly with the assigned regimen, **ITT** is helped little by an improvement in compliance. However, **ITT** is noticeably hurt by poor compliance.

5.8 Summary of Performance With Respect to Estimation Precision

One may interpret the results of studies 1-5 as supporting two different approaches to data analysis. One tactic is to develop guidelines based on study parameters. Depending on the parameters of a particular data set, the method that is expected to yield the best exposure-response estimates is selected. The second approach is to find one best analysis method across all possible designs and apply it generically. Both are outlined here.

Based on the plot of $\delta_{D=Dn}$ as a function of cor(C_Q,C) for study 1 (Figure 5.2), **ML** should be used if cor(C_Q,C)<.8. Based on the Δ_R boxplot for study 2 (Figure 5.5), **BSR** should be used if there is perfect correlation between C_Q and C. Study 2 also shows that it is inefficient to measure compliance using an electronic monitor if cor(C_Q,C)=1. Based on the plot of $\delta_{D=Dn}$ as a function of f_M for study 1 (Figure 5.3), **ML** should be used if fewer than 80 percent of subjects have a measure of C_M but **CD** should be used if f_M>.8. Based on the results of study 5 (Figure 5.10), **BSR** should be used if there are no C_M data at all. The recommendations seem disjointed since these guidelines are based on examining plots of error in P(Y|D=D_n) estimates as a function of, at most, one simulation variable at a time.

It is important to consider how simulation parameters simultaneously influence exposure-response estimates. For example, the optimal f_M for methods that use both self-reported and electronically monitored compliance data may differ depending on cor(C_Q,C). Contour plots representing the three-dimensional surface of $\delta_{D=Dn}$ as a function of two simulation parameters are used to illustrate the interaction between two simulation variables. To normalize precision to **ALL**, **ALL**'s $\delta_{D=Dn}$ for a given trial is subtracted from each method's $\delta_{D=Dn}$ and referred to as $\delta_{D=Dn}$' ($\delta_{D=Dn}' = \delta_{D=Dn}^{METHOD} - \delta_{D=Dn}^{ALL}$). Therefore, the plots are referred to as relative contour plots.

The clinical trial parameters of interest are plotted on the x- and y- axes and the corresponding $\delta_{D=Dn}$' value (in the z-axis) is indicated by points in the x,y plane. The magnitude of $\delta_{D=Dn}$' is represented by contour lines connecting points of equivalent estimation precision. The points of equivalent precision are determined by a smooth, or a local average, through the $\delta_{D=Dn}$' data. (The lowess function in Splus was used to smooth through the data.) The lines are drawn at equivalent precision intervals, so the spacing of the lines indicates the dependency of estimation precision on the variables. The closer the lines are, the stronger is the relationship between $\delta_{D=Dn}$' and the variable of interest.

Figure 5.22 represents the three dimensional surface of $\delta_{D=Dn}$' as a function of cor(C_Q,C) and f_M in study 1. Judged simply in terms of the magnitude of $\delta_{D=Dn}$', **ITT** is the poorest method of analysis over all regions of parameter space. In contrast, **ML** is the only method of analysis that yields the same value of $\delta_{D=Dn}$' as **ALL** at some region in parameter space ($\delta_{D=Dn}$' equals zero in the lower right hand corner of the figure).

Note that in study 1, N and ρ are varied in simulation in addition to f_M and cor(C_Q,C). Adjacent points in the cor(C_Q,C) and f_M plane may be far apart in the five-dimensional space that includes values of N and ρ. When there are many points in a region of the (f_M, cor(C_Q,C)) plane, the general trend between (f_M, cor(C_Q,C)) and $\delta_{D=Dn}$'

is observable, as differences due to variation in N and ρ average out. Figure 5.22 reveals that points are not evenly distributed across the cor(C_Q,C) axis—they are sparse at extreme values. Recall that *a* and *b*, the microparameters of P(C_Q|C), are Latin Hypercube Sampled from uniform distributions but cor(C_Q,C) is not uniformly distributed. This, most likely, explains any illogical waviness in the contour lines.

For example, **CD** does not use C_Q data to estimate P(Y|D=D_n), but its contour lines curl in to suggest that its precision decreases as cor(C_Q,C) increases. Additionally, **BSR**'s value of $\delta_{D=Dn}$ dips to a minimum when cor(C_Q,C) equals .6, but increases again for cor(C_Q,C) equal to .8. For this reason, general trends in the data are the focus. Taking a cue from **CD** and **BSR**, Figure 5.22 will only be trusted for cor(C_Q,C) ranging from .2 to .6.

Excluding edge effects, the contour lines for **BSR** run parallel to the f_M axis. This reflects **BSR**'s insensitivity to f_M. Likewise, **CD**'s contour lines run parallel to the cor(C_Q,C) axis. **ML**'s contour lines run parallel to the cor(C_Q,C) axis from cor(C_Q,C)=.2-.6—suggesting that its estimates are insensitive to cor(C_Q,C) in this range. At higher values of cor(C_Q,C), the lines run parallel to f_M, suggesting that there is no benefit incurred by adding C_M data when the amount of information in C_Q is high. However, the converse is not observed—the lines do not run parallel to f_M at low values of cor(C_Q,C). This indicates that **ML** is quite sensitive to f_M when there is little information in self-reported compliance. The contour lines fall in several directions on the plot for **ITT**, indicating that this method is insensitive to either parameter.

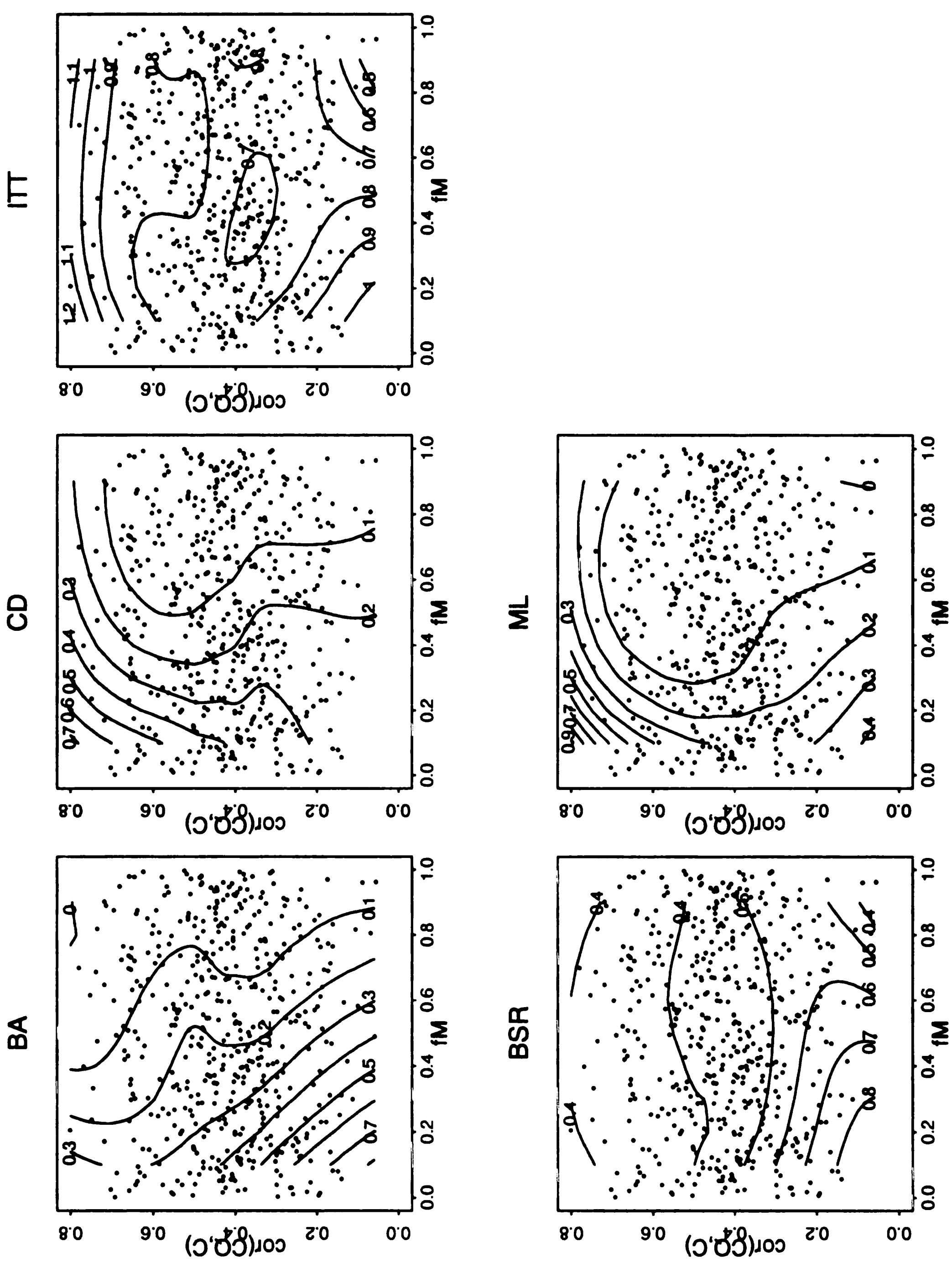


Figure 5.22. Contour Plot of δ_{D-D_M}' as a Function of cor(CQ,C) and fM in Study 1.

BA's contour lines cross the plotting region at, approximately, a 45 degree angle. This suggests that **BA**'s precision is improved to the same extent by equivalent fractional increases in cor(C_Q,C) and f_M. Therefore, when designing a trial that is to be analyzed using **BA**, one has the flexibility of improving estimates of exposure-response by altering whichever design feature is more feasible to change. In reality, f_M is likely the more adjustable parameter. The contour lines are most tightly spaced at low values of cor(C_Q,C) and f_M, revealing that the most improvement in estimation is to be gained when adding information in C_Q or adding C_M data when there is little to start with. **ML**'s lines are spaced further apart in the low f_M, low cor(C_Q,C) region than **BA**, indicating that it is much less sensitive to either parameter than **BA**.

Figure 5.22 can guide the selection of an analysis method based on values of f_M and cor(C_Q,C) in a data set. For example, consider a real data set taken from published studies(Burney, Krishnan et al. 1996; Straka, Fish et al. 1997) in which both self-reported compliance and electronically monitored compliance are measured in the same individual and data are reported in the body of the manuscript. The data from these two reports are pooled to yield a new data set with 85 joint values of C_M and C_Q. The measures of compliance, reported on a 0%-100% scale, require categorization to allow for interpretation with respect to contour plots. Reported percent compliance values are transformed into one of three categories by the following algorithm to yield a distribution such that $P(C_M=0)=P(C_M=.5)=P(C_M=1)=1/3$.

% Compliance	**Compliance Category**
<50	0
≥50-90	.5
≥90	1

Measured C_Q is not equivalent to C_M, so applying these cutoffs does not yield a uniform distribution of C_Q values.

This procedure for categorizing compliance data does not diminish the relationship between C_M and C_Q. The correlation between the two compliance measures for the pooled raw data set is .55, while correlation in the pooled categorized data set is .527. According to Figure 5.22, when cor(C_Q,C)=.527, **ML** yields estimates of drug effect with the least error for all values of f_M.

Figure 5.23, a relative contour plot representing the three-dimensional surface of $\delta_{D=Dn}$' as a function of (ρ,cor(C_Q,C)) in study 1, is a guideline for selecting an analysis method with respect to drug effect size and cor(C_Q,C) simultaneously. Of all analysis methods, **ITT**'s contour lines are the most parallel to the cor(C_Q,C) axis and the most closely spaced. This indicates that **ITT** is more strongly influenced by drug effect size than any other method. Interestingly, **BSR** is more strongly influenced by drug effect size than cor(C_Q,C). Likewise, ρ wields a greater influence on **BA** than cor(C_Q,C). However, **BA**'s relative contour lines are farther apart and smaller in magnitude than **BSR**'s $\delta_{D=Dn}$'—indicating that it is less sensitive to either parameter overall. Excluding edge effects, both **CD** and **ML** have widely spaced relative contour lines—indicating negligible dependency on either cor(C_Q,C) or ρ.

Study 5—an exploration of method performance in the extreme case where there are no C_M data—reveals interesting features of method behavior that are not apparent in study 1 results. Therefore, a relative contour plot representing the three dimensional surface of $\delta_{D=Dn}$' as a function of cor(C_Q,C) and ρ in study 5 is provided (see Figure 5.24). Note that aside from sampling variability, the plots of **BSR** and **ITT** are identical

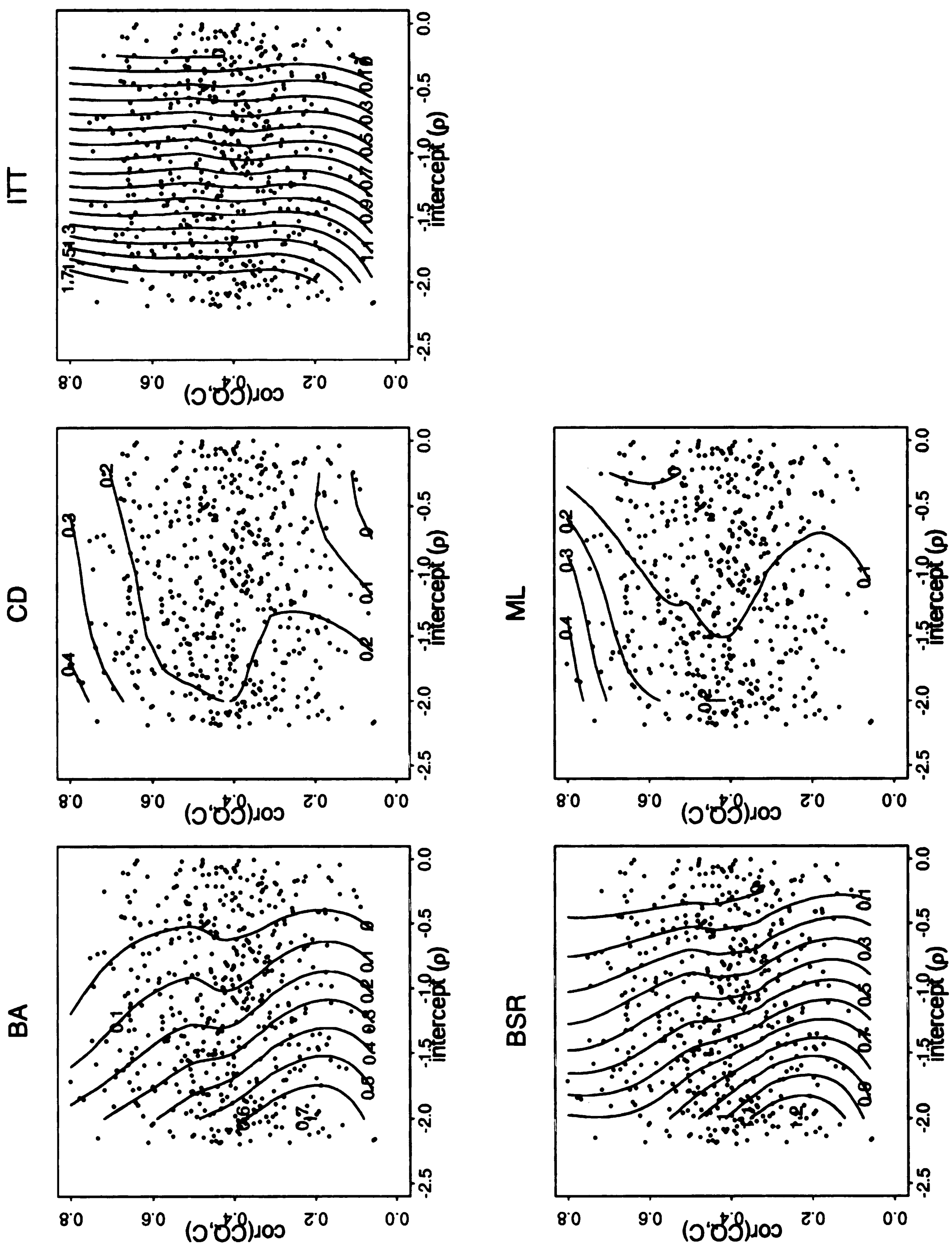


Figure 5.23. Contour Plot of δ_{D-Ds}' as a Function of cor(CQ,C) and ρ in Study 1.

to the corresponding study 1 plot (Figure 5.23) as neither method is influenced by a change in f_M. A comparison between **BSR**'s and **ITT**'s contour lines in Figure 5.24 to their respective lines in Figure 5.23 reveals that sampling variability "changes" method performance at the edge of the plots where cor(C_Q,C) data is sparse. In Figure 5.23, both **BSR**'s and **ITT**'s lines curl toward the cor(C_Q,C) axis at low values of cor(C_Q,C). In Figure 5.24, both **BSR**'s and **ITT**'s lines curl away from the cor(C_Q,C) axis in this region. These edge effects are to be ignored.

ML yields the least error in P(Y|D=D_n) estimates of all analysis methods. Its largest value of $\delta_{D=Dn}$' is 1.2, while **BA** and **BSR** yield values as high as 1.8 in Figure 5.24. Interestingly, as cor(C_Q,C) increases, **ML**'s contour lines bow away from the cor(C_Q,C) axis—indicating that for a given drug effect size, error in **ML** increases as cor(C_Q,C) increases. To some extent, this may be an artefact of edge effects in smoothing. However, by setting f_M=0, there is less variability in this data set than in Figure 5.22 and Figure 5.23, so this may reflect the influence of the prior P(C|C_Q) on **ML**. Recall that θ_{2prior} has no correlation between C_Q and C and **ML**'s estimates of drug effect worsen slightly as the data diverge from θ_{2prior}. Yet Figure 5.24 shows that the effect of the prior is moderate—**ML**'s performance is more greatly influenced by drug effect size.

This set of guidelines, albeit consistent with the results of simulation study, is difficult to apply in practice since it requires referring to charts or remembering arbitrary cutoff values. Futhermore, when planning a data analytic approach, some parameters, such as f_M and cor(C_Q,C), may be unknown to the investigator. A deeper flaw of this approach is that the exact values of these cutoffs are expected to be influenced by the trial design and data format. As such, this first approach can only be recommended for analysis of data of the same type investigated here. Because of its simplicity and because

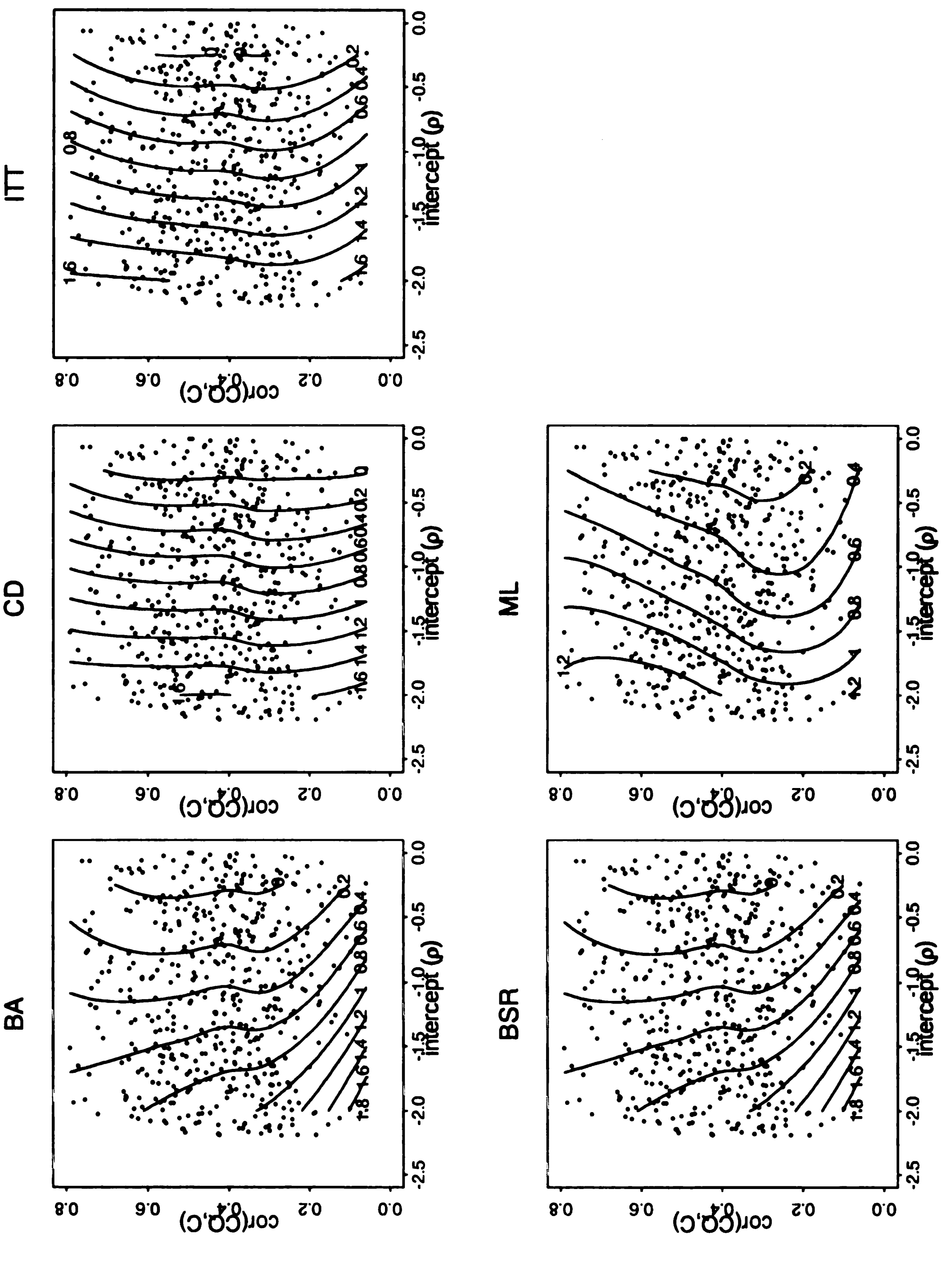


Figure 5.24. Contour Plot of δ_{D-D_m}' as a Function of cor(CQ,C) and ρ in Study 5.

it is more likely generalizable between data types, it is more attractive to find one method of analysis that yields the best exposure-response estimates across all study designs. If all of the results of studies 1-5 are taken as the evidence for this judgement, a procedure for deciding which method is best is needed.

One simple way to compare method performance across studies 1-5 is to tally up each method's successes with respect to rank order of Δ_R estimates. For example, the number of times each method has the smallest Δ_R reveals that both **ML** and **BA** are the best methods of analysis in three of the five studies, **BSR** is the best method of analysis in two of the five studies, and **CD** is the best method of analysis in one of the five studies. **ITT** is never the best method of analysis. **ML** outperforms **BA** in two of the five studies. Likewise, **BA** outperforms **ML** in two of the five studies. **ML** outperforms **BSR** in three of the five studies and outperforms **CD** in four of the five studies. **BA** outperforms **BSR** in three of the five studies, but only outperforms **CD** in two of the five studies.

By tallying rank ordered performance, **ML** is the best method of analysis, with a slight advantage over **BA**. Since both use C_Q and C_M data, the performance of **ML** relative to **BA** quantifies the gain in efficiency incurred by calibrating C_Q. By this estimation, it seems that the data analyst profits little from **ML**.

Tallying up performance in this way is misleading, however, for two reasons. First, rank ordering discards information about the magnitude of a method's advantage. Second, since studies 2-5 are carried out under a subset of conditions within study 1, it is unfair, perhaps, to weight studies 2-5 as heavily as study 1. Since no summary statistic adequately evaluates the advantage of any one method, the results of studies 1-5 are considered with respect to several statistics.

Taken together, the results suggest that the **ML** method is the best choice for analyzing calibration studies. **ML** is the most dependable. Its performance is the least sensitive to study design and uncontrollable factors such as drug effect size and the accuracy of C_Q. **ML** offers the most potential gain when there are missing C_M data, with no cost incurred when all subjects have a measure of C_M. **ML** performs remarkably well when there is little information in C_Q, with a small cost incurred if self-reported compliance is accurate.

Given the small advantage of **BSR** over **ML** in the extreme case that C_Q=C (refer to Figure 5.5) or f_M=0 (refer to Figure 5.10), relative to the cost of **BSR** if cor(C_Q,C)<1 and f_M>0 (refer to Figures 5.1, 5.7, and 5.9), **ML** is preferred to **BSR**. Furthermore, since one has no way of knowing if C_Q=C, diagnostic information for recommending **BSR** is unavailable. **ML** delivers its greatest payoff when fewer than 50% of subjects have C_M data—as evidenced by the distance between **ALL** and **ML** relative to the distance between **ALL** and all other methods in study 1 (Figure 5.3). In nearly all cases investigated, any method that uses compliance data is beneficial relative to the standard intention-to-treat procedure.

Although clinical trials run with f_M=0 aren't calibration designs, trials without C_M data allow one to determine how the **ML** method behaves in the generic situation in which it has nothing but prior information on P(C|C_Q). This has a number of practical applications. It allows one to evaluate method performance given the situation that no subject in a clinical trial has C_M data, but one wants to use prior information on P(C|C_Q) from the literature. A less obvious example is when one only has compliance information on one drug in a combination regimen (C_M^1) and the data analyst is not willing to assume that patients are equally compliant with other unmonitored drugs (C_M^2, C_M^3, etc.). Instead, one may use C_M^1 and C_Q data as prior information on the relationship between

C_M^2 and C_Q. The investigation reveals that this approach is reasonable as long as the prior $P(C|C_Q)$ is informative. **ML** is equivalent to **ITT** when there are no C_M data (see Figure 5.11) aside from the region in which the prior has good agreement with the true value of $P(C|C_Q)$. Thus, it is no worse than the standard method of analysis.

The results of studies run without a placebo group, studies run with the prior $P(C|C_Q)$ assuming $cor(C_Q,C)=1$, and studies with nonuniform $P(C)$ suggest that the benefit of **ML** is not an artifact of the assumption that $D(C,D_n=0)=0$, of the chosen value of θ_{2prior}, or of $P(C)$.

5.9 Power Under Conditions Favoring Particular Analysis Methods

5.9.1 Why Estimate Power?

Because the units of Δ_R lack intuitive meaning, power, or the probability of rejecting the null hypothesis of no drug effect when it should be rejected, is used to evaluate methods, in addition to estimation precision. Section A.3.2 of the Appendix gives a detailed description of how power is determined.

The power of each method is computed at several points in parameter space. Unlike in studies 1-5, parameter values are not chosen at random by Latin Hypercube Sampling from a range of values. In fact, the conditions of each study are chosen to favor one of the pseudo likelihood methods over all others. This serves to highlight how other methods measure up under extreme conditions. Knowledge gained in studies 1-5 guides the selection of the parameter values used.

Table 5.21 shows the fixed parameter values that are common among all power studies. Boldface type is used to indicate which fixed parameter values differ from study

1. P(Y|D=0) is fixed to .3 when data are simulated under the alternative hypothesis (H_a), and P(Y|D=0) is fixed to .5 when data are simulated under the null hypothesis (H_o).

5.9.2 ITT Favored Design—Study 6

A study in which a high percentage of patients actually comply with the prescribed regimen is expected to yield good performance by **ITT** because the **ITT** method assumes perfect compliance. For this reason, in Study 6, P(C) is selected to assign patients to C=1 with 50% probability and to each of C=0 and C=.5 with 25% probability. To assure that **ITT** performs better than the other methods, Study 6 is designed with few C_M data and poor information in C_Q—f_M is fixed to .15 and the weight of $C_{Q,Q\perp C}$ in constructing $P(C_Q|C)$ is 1. Table 5.22 summarizes these values.

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 5.23.

Figure 5.25, the boxplot of power computed using a χ^2 critical value under the conditions investigated in study 6, shows that, as expected, **ITT** has the greatest power of all methods with power = .74. Surprisingly, **ML** is a close competitor of **ITT** with power=.63. **BA** has 40% less power than **ITT** with power=.45. **BSR** and **CD** trail behind with power equal to .36 and .19, respectively.

Figure 5.26, the boxplot of Δ_R under the conditions investigated in study 6, shows that the rank order of power is consistent with the rank order of Δ_R for all methods but **ITT**. By Δ_R, **ML** is the best method of analysis, **BA** is second best, and **ITT** is third best.

If **ITT** yielded the most precise estimate of P(Y|D), the rank order of method performance would be consistent between Figures 5.25 and 5.26. This discrepancy in **ITT**'s performance is understandable given a more in depth consideration of power.

Table 5.21. List of Fixed Parameter Values Common to All Power Studies.

Parameter	**Fixed Value**
λ_1	.5/3
λ_2	.5/6
θ_{1prior}	$P(Y\|D=0)=P(Y\|D=.5D_n)=P(Y\|D=D_n) = .5$
θ_{2prior}	$C_{Q,Q\perp C}$
$P(C_M\|C)$	$C_{M,M=C}$
N	**100**
ρ under H_a	**logit(.3)**
ρ under H_o	**logit(.5)**

Table 5.22. Distribution of Random Parameter Values in Study 6.

Parameter	**Fixed Value**
f_M	.15
$WT_{Q=C}$	0
$WT_{Q\geq C}$	0
$WT_{Q\perp C}$	1
P(C)	P(C=0)=.25, P(C=.5)=.25, P(C=1)=.5

Table 5.23. Observed Range of Parameter Values in Study 6.

Parameter	**Range**
f_M	.15
cor(C_Q,C)	-.197—.224
P(C)	P(C=0): .16-.36, P(C=.5): .18-.35, P(C=1): .37-.63

Power and Δ_R reflect different aspects of method performance. Power, or the ability of methods to discriminate between estimates of P(Y|D=0) and P(Y|D=D_n), is sensitive to variance. It should be high for methods that yield precise estimates of P(Y|D=0) and P(Y|D=D_n). As a measure of error, Δ_R reflects a more equitable sensitivity to both bias and variability.

The discrepancy between **ITT**'s performance with respect to Δ_R relative to power suggests one limitation of **ITT**. It can answer a yes/no question (*Is the mean response in the D=0 and D=D_n groups significantly different?*) adequately, however, it cannot be trusted to deliver an estimate of that difference. The preferred metric depends on what is of interest to the investigator.

A comparison of Figure 5.27, the boxplot of power computed using a simulated critical value, to Figure 5.25 shows that **CD** is the only method influenced by the source of the critical value. All methods but **CD** have the same power regardless of whether the critical value is taken from a χ^2 table or simulated under H_o. (Refer to the Appendix for more detail.) **CD** performs better using a simulated critical value than using the χ^2 value. Closer inspection reveals that **CD**'s simulated critical value is similar to the χ^2 value

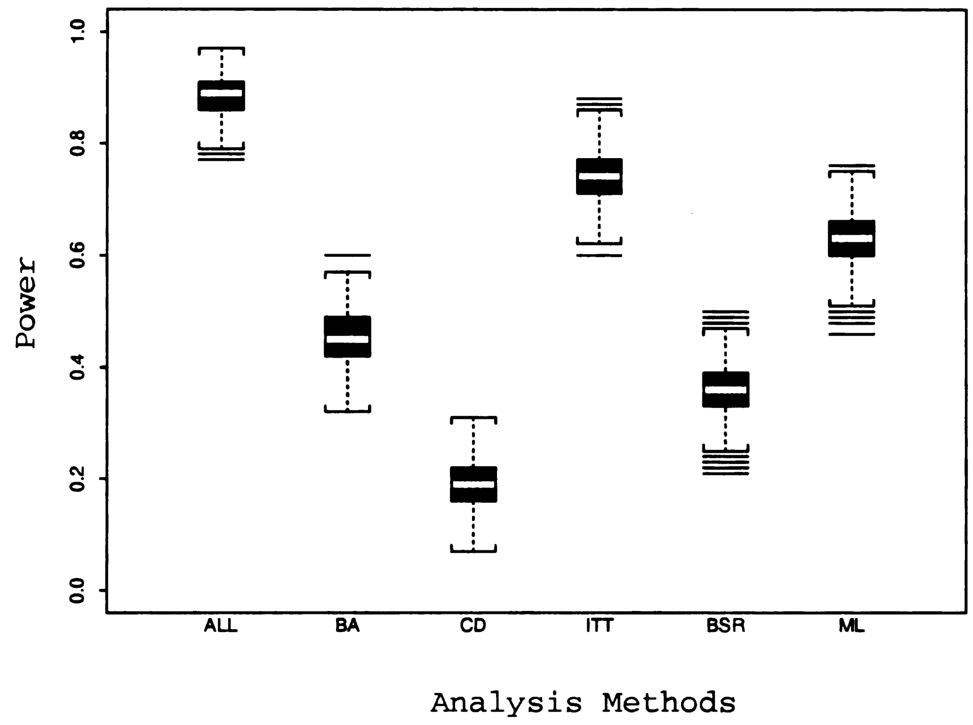


Figure 5.25. Boxplot of Power for Study 6 (χ^2 Critical Value). Good compliance, few C_M data, and poor agreement between C_Q and C favor the **ITT** method with respect to power. **ML** is the next most powerful method.

given 1 degree of freedom rather than 2. This makes sense considering the high probability that there is no data available for **CD** to estimate P(Y|D=.5) or P(Y|D=1) under this study design. That is, of the 15 subjects with C_M data (f_M=.15 * N=100), half have D_n=1. Of these 7.5 subjects, there is a .25 probability of having C=.5 and a .5 probability of having C=1. Thus, on average, in any given trial, only 1.875 and 3.75 subjects have D=.5 and D=1, respectively. Given the luck of the draw, no subject may have D=.5 or D=1. In that case, **CD** fills in the prior on P(Y|D=.5) or P(Y|D=1). Thus, analogous to the explanation of why **ITT** uses a χ^2 critical value with only one degree of freedom (refer to the Appendix), **CD** should be evaluated under such conditions using a χ^2 critical value with only one degree of freedom.

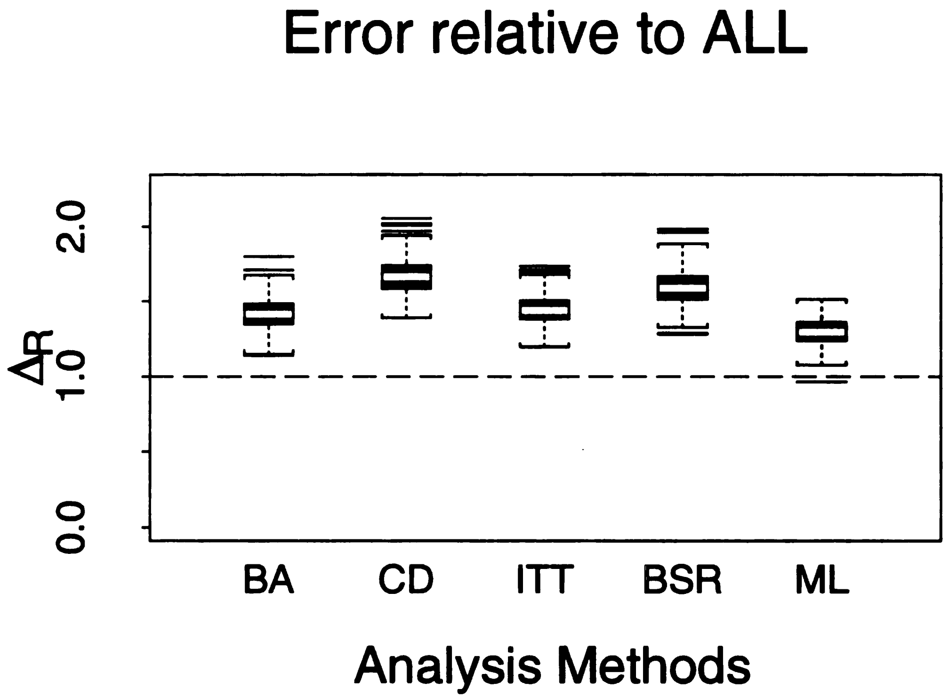


Figure 5.26. Boxplot of Δ_R for Study 6. Contrary to performance as measured by power, in study 6, **ML** yields estimates of drug effect with the lowest error. **ITT** yields the third most precise estimates of P(Y|D).

Regardless, relative method performance is not changed by this improvement in **CD**. **CD**, along with **BSR**, has the lowest power and yields the poorest estimates of P(Y|D) under this design. However, this result supports the use of the χ^2 critical value as long as the degrees of freedom are chosen correctly.

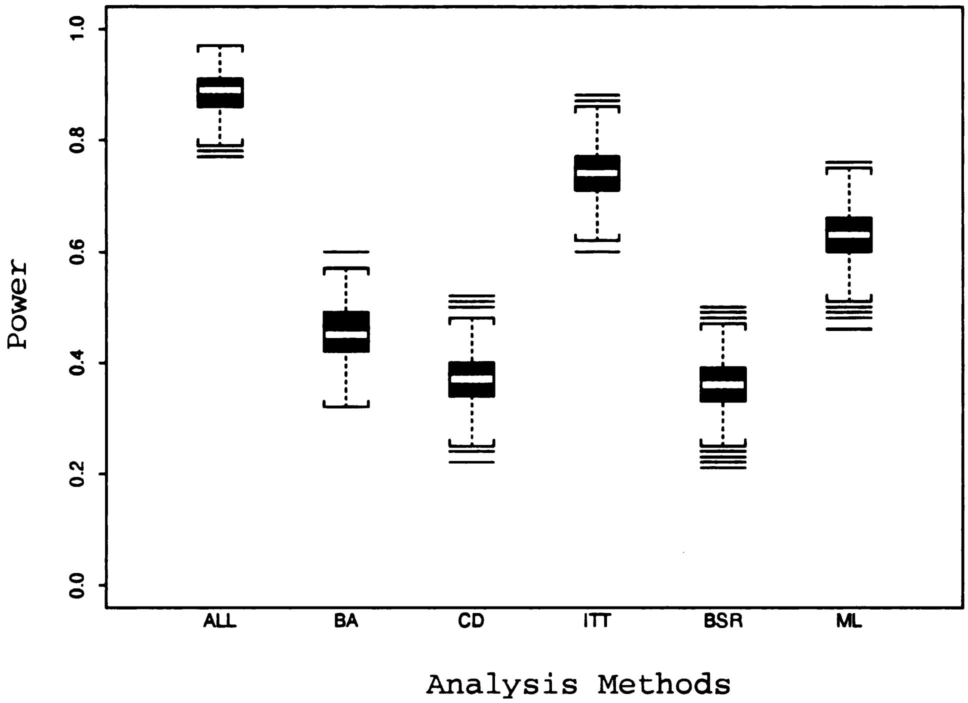


Figure 5.27. Boxplot of Power for Study 6 (Critical Value Determined by Simulating Under H_o). Compared to Figure 5.25 in which power is determined using the χ^2 critical value, all methods but **CD** have the same performance. The discrepancy in **CD**'s performance reflects the influence of study design on the number of degrees of freedom for **CD**. The similarity between Figure 5.25 and Figure 5.27 suggests that the χ^2 critical value is appropriate for computing power, provided that the appropriate degrees of freedom are chosen for **CD**.

5.9.3 CD Favored Design—Study 7

A study in which a high percentage of patients have C_M data, there is poor correlation between C_Q and C, and the distribution of P(C) is skewed toward low intake is expected to show **CD** in its best light. For this reason, in Study 7, f_M is set to .6, the weight of $C_{Q,Q\perp C}$ is fixed to 1, and P(C) assigns 50 percent of patients to C=0 and 25 percent of patients to each of C=.5 and C=1. Table 5.24 summarizes these values.

Table 5.24. Distribution of Random Parameter Values in Study 7.

Parameter	**Fixed Value**
f_M	.6
$WT_{Q=C}$	0
$WT_{Q\geq C}$	0
$WT_{Q\perp C}$	1
P(C)	P(C=0)=.5, P(C=.5)=.25, P(C=1)=.25

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 5.25.

Table 5.25. Observed Range of Parameter Values in Study 7.

Parameter	**Range**
f_M	.6
cor(C_Q,C)	-.207—.211
P(C)	P(C=0): .37-.63, P(C=.5): .16-.33, P(C=1): .14-.39

Figure 5.28, the boxplot of power for each method in study 7 computed using a χ^2 critical value, shows that, as expected, **CD** performs well. However, **ML** has as much

power as **CD**. A comparison of Figure 5.29, the boxplot of power computed using a simulated critical value, to Figure 5.28 shows that **CD** is influenced by the source of the critical value. Analogous to the result presented in Figure 5.27, the influence of study design on **CD**'s degrees of freedom is at fault. **CD** has the most power of all methods.

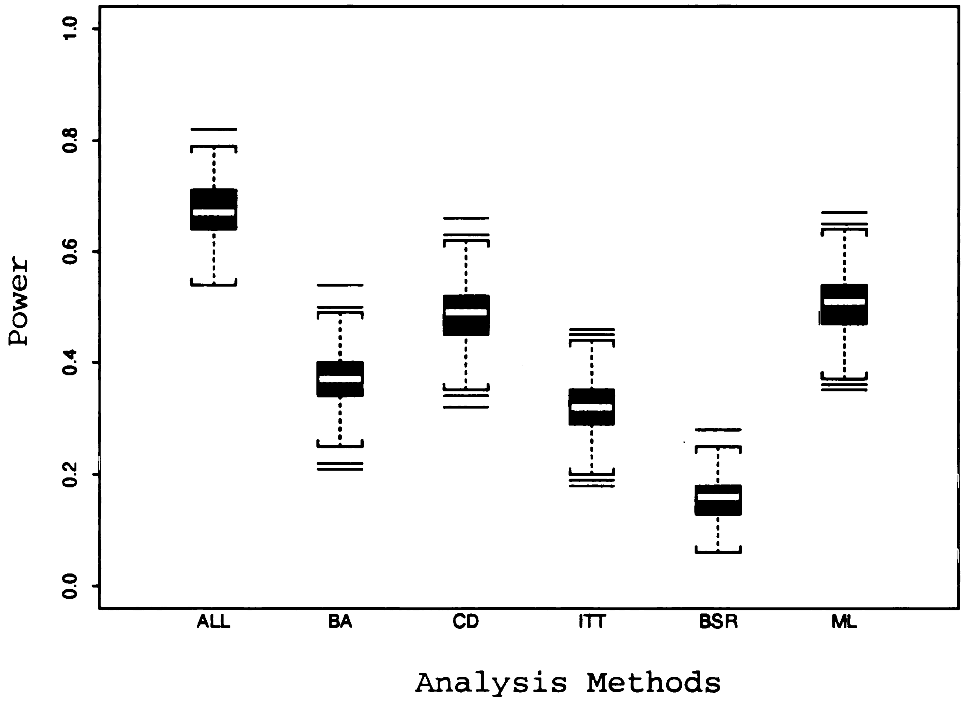


Figure 5.28. Boxplot of Power for Study 7(χ^2 Critical Value). Poor compliance, rich C_M data, and poor agreement between C_Q and C favors **CD** and **ML** with respect to power.

Under the conditions investigated in study 7, the rank order of method performance with respect to power is different than performance with respect to Δ_R (compare Figure 5.30 to Figure 5.29). To be specific, **BA**, **CD**, and **ML** yield equivalent

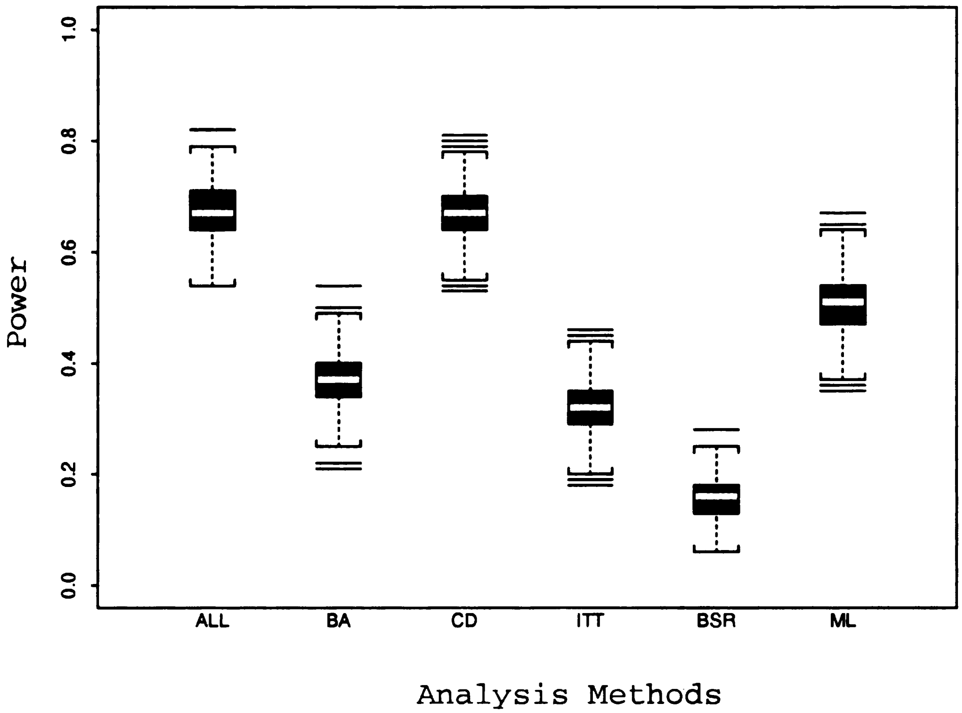


Figure 5.29. Boxplot of Power for Study 7 (Critical Value Determined by Simulating Under H₀). Compared to Figure 5.28 in which power is determined using the χ^2 critical value, all methods but **CD** have the same performance. The discrepancy in **CD's** performance reflects the influence of study design on **CD's** degrees of freedom.

estimates of Δ_R, but only **CD** and **ML** yield equivalent estimates of power. **ITT** and **BSR** yield equivalent estimates of Δ_R, but **ITT** has more power than **BSR**.

As in study 6, the dramatic shift between **ITT**'s performance with respect to power versus **ITT**'s performance with respect to Δ_R is consistent with **ITT** being a method that yields precise, but biased, estimates of P(Y|D). Conversely, the discrepancy between **BA**'s low power and **BA**'s high estimation precision reflects a high variability (apparent in the wide distribution of Δ_R for **BA**), but low bias (indicated by its mean Δ_R

close to 1) in θ_1 estimation. **BA**'s Δ_R represents a tradeoff between the error incurred by using biased C_Q data relative to the increased precision gained by increasing the amount of data available.

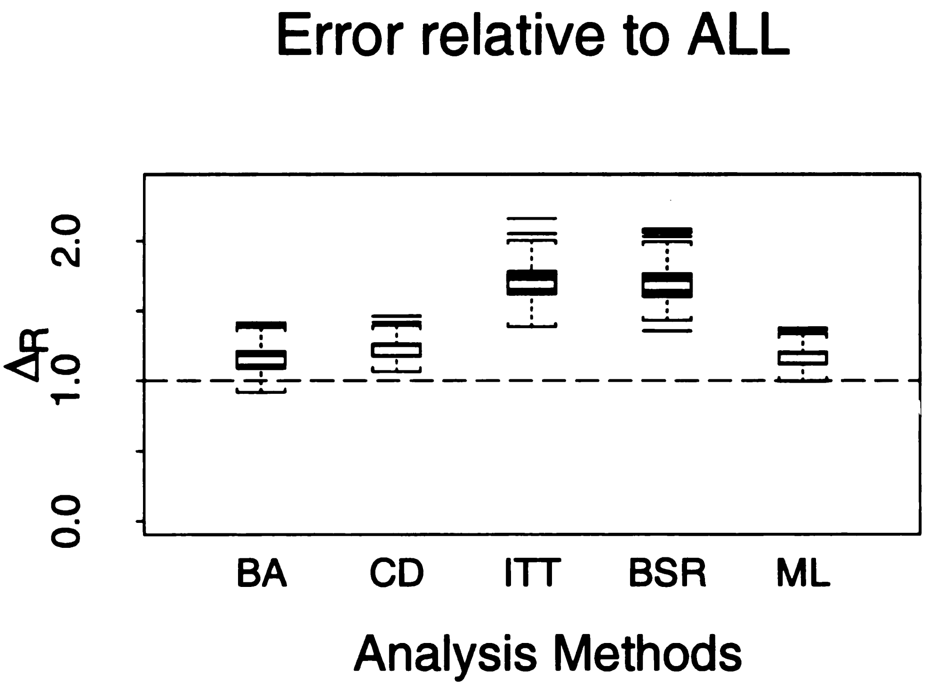


Figure 5.30. Boxplot of Δ_R for Study 7. Performance with respect to error in P(Y|D) estimates yields a different rank ordering of methods than performance as measured by power. **ITT** has more power than **BSR**, but yields P(Y|D) estimates that are as imprecise as **BSR**. **BA** has less power than **CD** and **ML** despite having P(Y|D) estimates that are as precise as **CD** and **ML**.

5.9.4 BSR Favored Design—Study 8

To favor **BSR**, a study with good correlation between C_Q and C but few C_M data is carried out. Table 5.26 summarizes the parameter values explored. Observed values of certain parameters in the data actually simulated in a typical study are given in Table 5.27.

Table 5.26. Distribution of Random Parameter Values in Study 8.

Parameter	Fixed Value
f_M	.1
$WT_{Q=C}$	.2
$WT_{Q\geq C}$	.6
$WT_{Q\perp C}$	.2
P(C)	P(C=0)=P(C=.5)=P(C=1)=1/3

Table 5.27. Observed Range of Parameter Values in Study 8.

Parameter	Range
f_M	.1
cor(C_Q,C)	.285—.678
P(C)	P(C=0): .23-.49, P(C=.5): .25-.46, P(C=1): .23-.47

Figure 5.31, a boxplot of power using the χ^2 critical value under the conditions investigated in study 8, shows that parameters favoring the power of **BSR** also favor

many other methods. **BA**, **ITT**, **BSR**, and **ML** yield equivalent estimates of power. Only **CD** has less power than **BSR**. **BA** has more power than **BSR** due to the small amount of C_M data available.

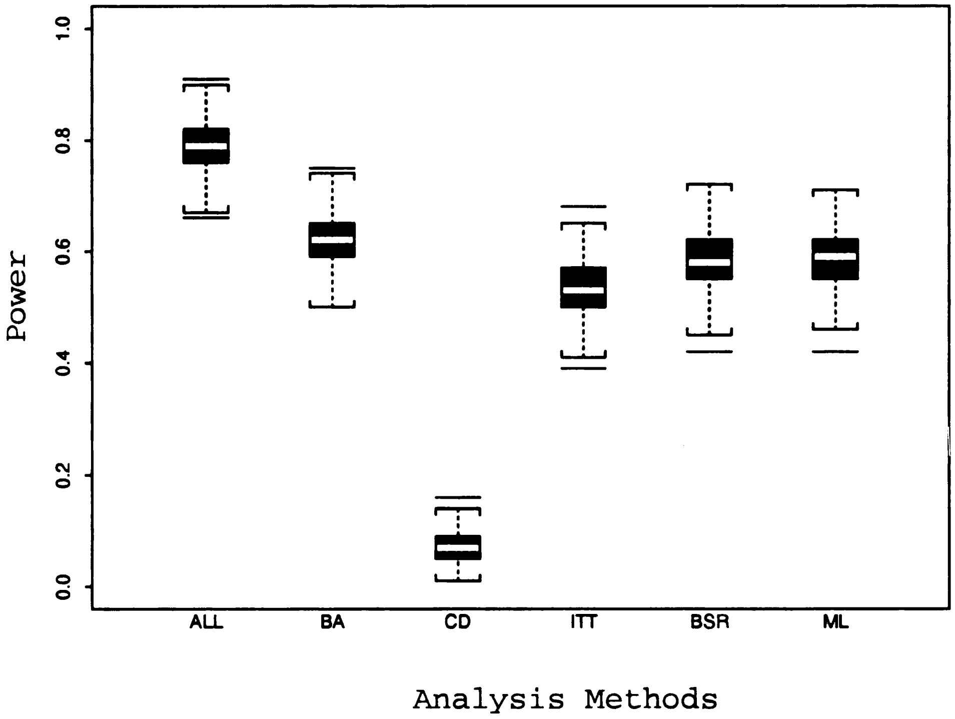


Figure 5.31. Boxplot of Power for Study 8 (χ^2 Critical Value). High correlation between C_Q and C and few C_M data favors the **BSR** method with respect to power. **BA**, **ITT**, and **ML** have as much power as **BSR** under these conditions.

Figure 5.32 shows that **CD**'s power changes slightly when the critical value is simulated relative to when it is computed using the χ^2 critical value. Contrary to the power shift observed in studies 6 and 7, this change is less dramatic. The result reflects the influence of the P(C) used in simulation. That is, by assigning an equivalent number of subjects to C=0, C=.5, and C=1, it is less likely that **CD** doesn't have sufficient data to estimate a component of P(Y|D). Relative method performance is unaffected by the source of the critical value so this result is not considered futher.

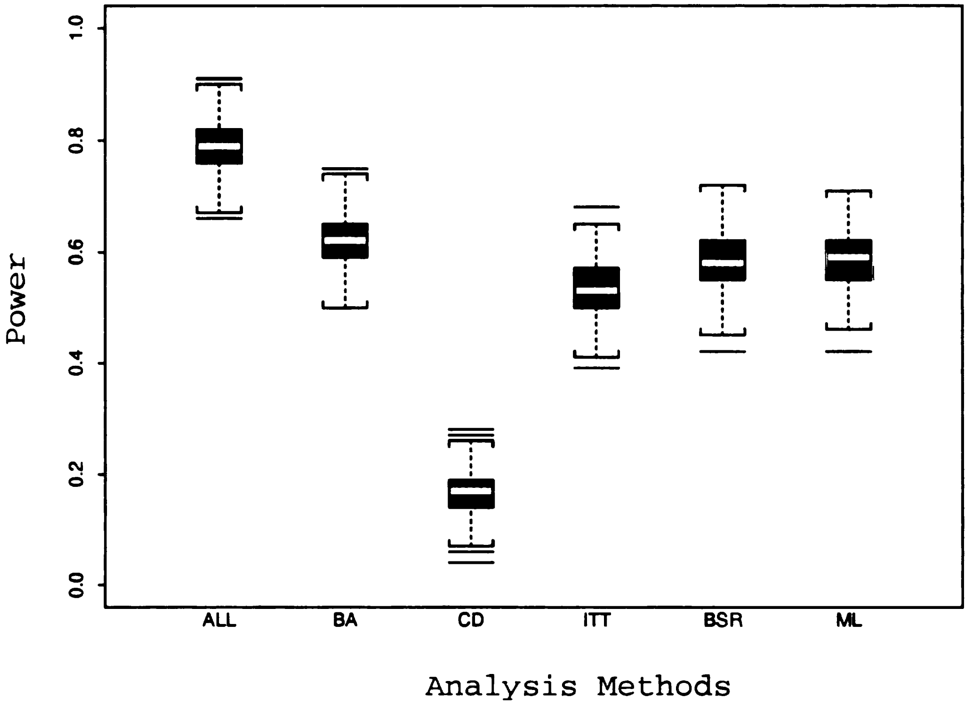


Figure 5.32. Boxplot of Power for Study 8 (Critical Value Determined by Simulating Under H_o). Compared to Figure 5.31 in which power is determined using the χ^2 critical value, all methods but **CD** have the same performance. The discrepancy in **CD**'s performance reflects the influence of study design on **CD**'s degrees of freedom.

Figure 5.33, a boxplot of Δ_R for study 8, shows that the rank order of methods with respect to power is consistent with Δ_R for all but **ITT**. As in study 6 and study 7, Δ_R for **ITT** is due mostly to bias. But **ITT** has high power because there is little variability in its estimates. In contrast, **CD** has very low power because there is high variability in its estimates—a consequence of having few C_M data available to estimate P(Y|D).

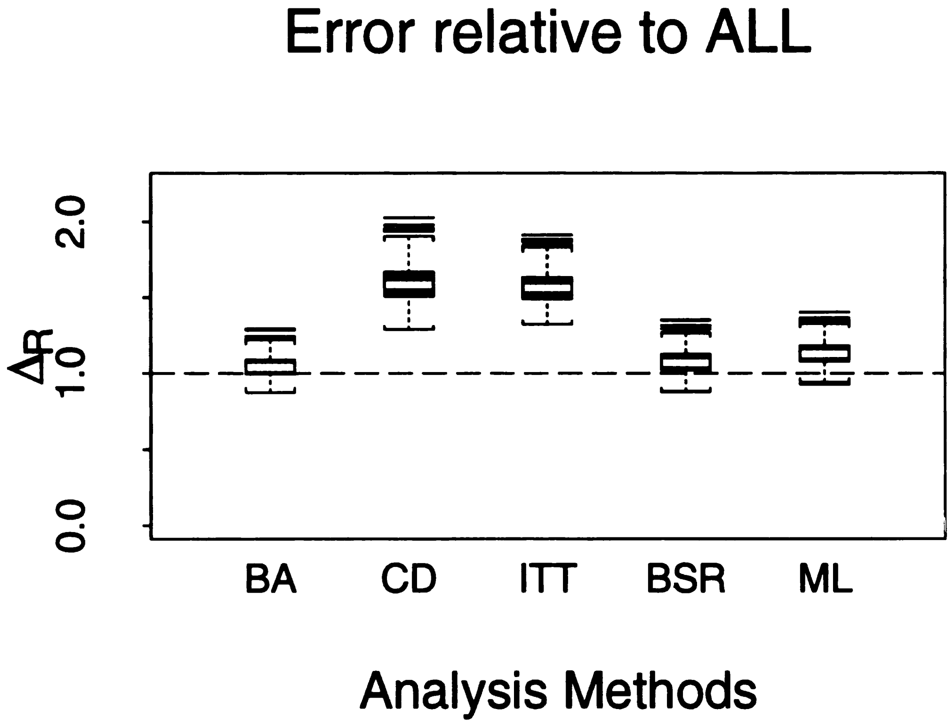


Figure 5.33. Boxplot of Δ_R for Study 8. With the exception of ITT, performance with respect to error in P(YID) estimates is consistent with performance as measured by power.

5.9.5 Summary of Performance with Respect to Power

Power often guides clinical trial (and experimental) design. Of the factors that affect power—number of subjects investigated (N), effect size, variability in effect, and level of desired statistical significance—N is the only one commonly perceived as under the investigator's control. Another, often overlooked, factor that contributes to power and is under the control of the experimentalist is the method of data analysis. This work illustrates that power can be improved by paying more attention to data analysis—an option that may be more economical than increasing trial enrollment and the only decision that remains flexible after a trial's completion.

The results of the power studies offer further support in favor of using the **ML** method. Although there are methods preferred to **ML** in special circumstances, it is consistently among the top performers. In studies 6, 7, and 8, **ML**'s power falls within a tight range—from .51 to .63. In contrast, **BA**'s power ranges from .37 to .62, **CD**'s power ranges from .17 to .66 (power as determined using a simulated critical value), **ITT**'s power ranges from .32 to .74, and **BSR**'s power ranges from .16 to .58. The results of these studies are all the more compelling considering that none of them were designed to favor **ML**.

5.10 Discussion: Method Performance Under Ideal Conditions

In this chapter, method performance is evaluated over widely varying clinical trial designs, patient compliance distributions, and drug effect sizes. Studies 1-5 suggest that **ML** is overall the best method of analysis across the parameter space explored, while **ITT** yields the poorest estimates of P(Y|D). Studies 6-8 demonstrate that **ML** is consistently among the most powerful methods of analysis, but surprisingly reveal that

ITT has more power than one might expect based on its performance with respect to Δ_R. This raises an important question. Do studies 1-5 unfairly represent **ITT**'s performance?

After all, the distribution of true compliance (of interest because **ITT** is most sensitive to P(C)) in study 8—where **ITT** is among the most powerful methods of analysis—is the same as in studies 1-5. The fixed values of N and ρ in study 8 fall in the middle of the range of randomly chosen values in studies 1-5. Therefore, study 8 can be thought of as providing an estimate of the average power of **ITT** across studies 1-5.

Whether the recommendations based on studies 1-5 should be amended depends on what is of interest to the data analyst. If one's goal is to reject the hypothesis that drug has no effect—as when satisfying the requirements of a regulatory agency—then as long as P(C) isn't skewed toward poor compliance (as in study 7), **ITT** may perform as well as the contending analysis methods. If it is important to estimate the actual exposure-response relationship, then **ITT** is not the method of choice. But given that P(C) is unmeasurable, **ITT** is ill advised. Since **BA**'s power is sensitive to cor(C_Q,C)—an unmeasurable parameter—this method should be used cautiously, as well.

Regardless of whether neither goal, power or unbiased estimation, is clearly dominant, it is important to understand a limitation of power. Power reflects a method's ability to yield distinct estimates—it rewards low variability in estimation more than bias. It is troubling since there are ways to decrease variability in drug effect estimates, and, consequently, increase power, that do not affect a difference in a drug's efficacy.

These results reflect method performance when there is no confounding between compliance and response. Confounding is not addressed in this report as it is a deep issue that has been explored elsewhere(Efron and Feldman 1991; Sheiner and Rubin 1995;

Angrist, Imbens et al. 1996; Goetghebeur, Molenberghs et al. 1998; Robins 1998). Furthermore, confounding would only need to be addressed here if there were some reason to believe that the confounding of drug response disproportionately affects data analysis methods. We have no reason to believe that one method for determining compliance might be more liable to co-vary with a confounder than another.

Note that studies 1-8 reflect method performance when ***A1-A5*** are satisfied. Thus, this chapter has presented **ML** in its best light. **ML**'s performance encourages the pursuit of a sensitivity analysis.

5.11 References

Angrist, J. D., G. W. Imbens and D. R. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables." Journal of the American Statistical Association **91**: 444-472.

Burney, K. D., K. Krishnan, M. T. Ruffin, D. Zhang and D. E. Brenner (1996). "Adherence to Single Daily Dose of Aspirin in a Chemoprevention Trial. An Evaluation of Self-Report and Microelectronic Monitoring." Archives of Family Medicine **5**(5): 297-300.

Caron, H. S. (1985). "Compliance: The Case for Objective Measurement." Journal of Hypertension **3**: 11-17.

Efron, B. and D. Feldman (1991). "Compliance as an Explanatory Variable in Clinical Trials." Journal of the American Statistical Association **86**: 9-22.

Goetghebeur, E., G. Molenberghs and J. Katz (1998). "Causal Effect of Compliance on Binary Outcome in Randomized Controlled Trials." Statistics in Medicine **17**: 341-355.

Robins, J. M. (1998). "Correction for Non-Compliance in Equivalence Trials." Statistics in Medicine **17**: 269-302.

Sheiner, L. B. and D. B. Rubin (1995). "Intention-to-Treat Analysis and the Goals of Clinical Trials." Clinical Pharmacology and Therapeutics **57**(1): 6-15.

Straka, R. J., J. T. Fish, S. R. Benson and J. T. Suh (1997). "Patient Self-Reporting of Compliance Does Not Correspond with Electronic Monitoring: An Evaluation Using Isosorbide Dinitrate as a Model Drug." Pharmacotherapy **17**(1): 126-132.

Chapter 6: Sensitivity Analysis

Abstract

In this chapter, sensitivity to assumptions regarding the accuracy of compliance measuring tools is explored. The results of this chapter show that **ML** is robust to violation of ***A3***—**ML** yields the best estimates of exposure-response when both C_M and C_Q have less than perfect correlation with C. **ML** is sensitive to ***A2***, but under the most realistic conditions tested, **ML** is equivalent to the best performing methods. When ***A2*** and ***A3*** are violated simultaneously, **BSR** yields the best exposure-response estimates, while **ML** and **BA** have the next best performance. In reality, ***A3*** is of great concern while ***A2***—the assumption that the accuracy of self-reported compliance is independent of the presence of an electronic monitor—is reasonable. Since good estimates of exposure-response are obtained using **ML** even if an electronic monitor does not measure true intake, the results suggest that there will be an improvement in **ML**'s performance if the likelihood is altered to allow for violation of ***A3***.

6.1 Purpose

The simulation studies carried out thus far reflect method performance under several assumptions (refer to Chapter 4 for an explanation of ***A1-A5***). Assumptions 2-5 are untestable and may not hold for real data sets. Therefore, a sensitivity analysis is performed to determine the extent to which the reported results depend on assumptions related to compliance measurement. Specifically, sensitivity to ***A2*** (random assignment to M such that $P(C|C_Q)$ is not influenced by an electronic monitor) and ***A3*** (the electronic monitor measures C) is determined. Robustness to *A4* and *A5* is not explored, as confounding is a complex issue, beyond the scope of this study focused on missing data problems. For further references to the confounding problem refer to the following references (Efron and Feldman 1991; Sheiner and Rubin 1995; Angrist, Imbens et al. 1996; Goetghebeur, Molenberghs et al. 1998; Robins 1998).

Since ML is the most assumption-laden method of analysis, its performance is predicted to be most affected in these studies.

6.2 Methods

Data are simulated using parameter distributions violating the assumption(s) of interest, then analyzed by all contending methods. The extent to which method performance changes relative to studies in which data are simulated in agreement with that assumption reveals its sensitivity to the assumption in question. To allow for direct comparison with the results of studies 1-5 (reported in chapter 5), Latin Hypercube Sampling of identical parameter values is performed (except for changes to the values in question). The following parameter values are common to all sensitivity studies. Note that Table 6.1 is identical to Table 5.1, except that it does not specify a fixed value for $P(C_M|C)$.

6.3 Investigation of Sensitivity to *A3* (C_M=C)

For all studies presented in chapter 5, C_M is simulated equivalent to C. In reality, electronically monitored compliance may not be an accurate measure of true intake. Some patients report removing doses for later ingestion along with the dose taken

Table 6.1. List of Fixed Parameter Values Common to All Sensitivity Studies.

Parameter	**Fixed Value**
N_{lhs}	100, ×5
P(C)	P(C=0)=P(C=.5)=P(C=1) = 1/3
λ_1	.5/3
λ_2	.5/6
θ_{1prior}	$P(Y\|D=0)=P(Y\|D=.5D_n)=P(Y\|D=D_n) = .5$
θ_{2prior}	$P(C=k\|C_Q=j)=1/3$ for j,k {0,.5,1}

immediately when it is more convenient to medicate in this way(Bangsberg, Hecht et al. 2000; Turner and Hecht 2001). If those doses are later taken without opening the electronically monitored bottle at the time of ingestion, the dosing event goes unrecorded and compliance is underestimated by C_M. To investigate sensitivity to error in C_M, the relationship between simulated C_M and C is changed from C_M=C to $C_M \leq C$.

6.3.1 Investigation of Performance Over a Wide Range of Parameter Space: C_M Underestimates C—Study $1^{CM\leq C}$

To determine robustness to *A3* over a wide range of parameter values, study 1 is repeated with one exception—C_M is not fixed equal to C. Table 6.2 shows the conditional distribution, $P(C_M|C)$, used to generate C_M data. Note that the similarity between $P(C_M|C)$ and $P(C_Q|C)$ as defined by $C_{Q,Q\geq C}$ (see Table A.1) indicates that C_M underestimates C by as much as C_Q overestimates it in $C_{Q,Q\geq C}$. For example, a patient with C=.5 has a fifty

Table 6.2. P(C_M|C) of Simulation for Sensitivity Studies Violating *A3*.

$C_{M,M\leq C}$		C		
		0	.5	1
C_M	0	1	1/2	1/3
	.5	0	1/2	1/3
	1	0	0	1/3

percent chance of having C_M=.5 and a fifty percent chance of underestimating compliance with C_M=0. The same patient has a fifty percent chance of self-reporting compliance accurately (C_Q=.5) and an equal chance of overestimating it with C_Q=1. In contrast to P(C_Q|C) of simulation, P(C_M|C) is not created by forming a linear combination of three prototypical distributions for the relationship between measured and true compliance. Only $C_{M,M\leq C}$ is used to simulate C_M data.

Table 6.3 specifies the range of parameters sampled for this study. Boldface type is used to indicate values differing from study 1 (compare Table 6.3 to Table 5.2).

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 6.4. (The range of cor(C_M,C) is given in boldface type to highlight the difference from Table 5.3.)

A comparison of the Δ_R boxplot for study $1^{CM\leq C}$ (Figure 6.1) to the Δ_R plot for study 1 (Figure 5.1) shows that **CD** is least robust to a violation of ***A3***. By this metric, all methods but **CD** maintain the same rank order of performance. **CD** yields poorer estimates of P(Y|D) when $C_M\leq C$ because, in addition to the imprecision caused by

Table 6.3. Distribution of Random Parameter Values in Study $1^{CM\leq C}$.

Parameter	Distribution
N	U(50,400)
f_M	U(0,1)
a, *b*	U(0,1), U(0,2)
ρ	U(logit(.1),logit(.5))
P(C_M|C)	**$C_{M,M\leq C}$**

Table 6.4. Observed Range of Parameter Values in Study $1^{CM\leq C}$.

Parameter	Range
N	50—400
f_M	.00273—.997
cor(C_Q,C)	.0546—.801
cor(C_M,C)	**.393—.711**
P(Y|D=0)	.100—.500
P(Y|D=D_n)	.500—.900
P(C)	P(C=0): .2—.5, P(C=.5): .188—.441, P(C=1): .154—.471

discarding data, its ability to estimate exposure-response is hurt by bias in C_M. Underestimating compliance has the same effect on estimation error as overestimating intake.

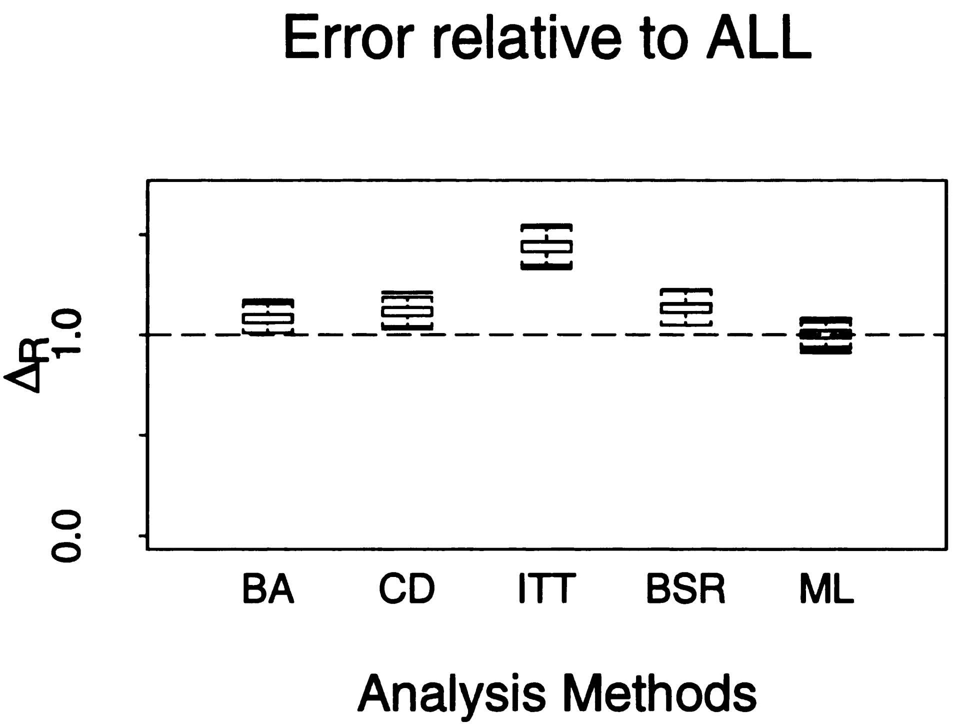


Figure 6.1. Boxplot of Δ_R for Study $1^{CM\leq C}$. When electronically monitored compliance underestimates true compliance, **ML** estimates exposure-response as well as **ALL** and better than the contending methods of analysis. The apparent improvement in **BSR** and **ITT** relative to study 1 (Figure 5.1) reflects a worsening of **ALL**'s P(Y|D) estimates.

Recall that **ALL** uses C_M, not C, data.Therefore, its exposure-response estimates suffer when $C_M \leq C$. Since Δ_R summarizes error in exposure-response estimates relative to **ALL**, a worsening of **ALL**'s performance is manifest as an improvement in Δ_R for **ITT** and **BSR**. Naturally, **ITT** and **BSR** are not influenced by a violation of *A3* since neither uses C_M data.

Figure 6.1 suggests that **ML** is robust to *A3*. In study $1^{CM\leq C}$, it performs better than all contending analysis methods. Interestingly, **ML** yields P(Y|D) estimates with nearly as much precision as **ALL** when *A3* is violated. This has an important implication for clinical trial design—it may be inefficient to measure compliance with an electronic monitor in all subjects when $C_M \leq C$.

Although **BA** is expected to approach **BSR** as cor(C_M,C) decreases to cor(C_Q,C), **BA**'s Δ_R does not lie between **CD** and **BSR**. The plot of $\delta_{D=Dn}$ as a function of f_M in Figure 6.2 shows why—the relative performance of **BA, CD**, and **BSR** depends on f_M.

BA is actually only better than **CD** when fewer than 40 percent of subjects have C_M data. As shown by cor(C_M,C) and cor(C_Q,C) in Table 6.4, C_M is a better measure of intake than C_Q. Although C_M is a better measure of C than C_Q, bias in C_M becomes a serious problem for **CD** at low values of f_M—**BSR** estimates exposure-response better than **CD** when f_M is less than .25. There is a point (here: f_M=.4) at which error in the estimate of P(Y|D) incurred by using biased, but abundant, C_Q data is greater than the error due to using a smaller data set with biased C_M data. Both cor(C_M,C) and cor(C_Q,C) are expected to drive the cutoff at which **BA, CD**, and **BSR** are favored relative to one another.

The plot of $\delta_{D=0}$ as a function of f_M in Figure 6.2 illustrates how underestimating compliance has the same effect on the estimate of P(Y|D) as overestimating intake. Here, **ALL** yields the least precise estimates of P(Y|D=0) for all values of f_M. Methods that do

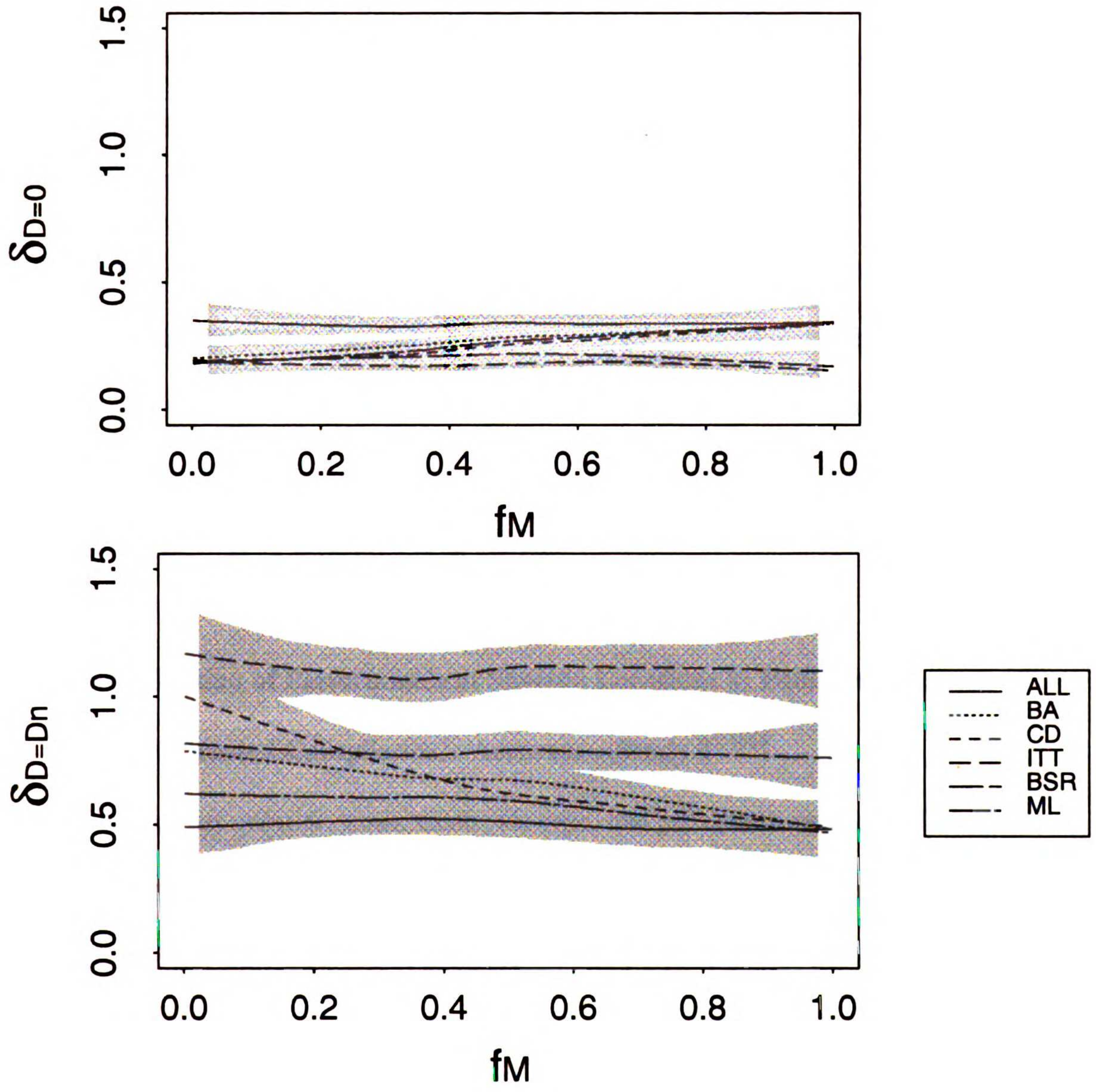


Figure 6.2. Plots of $\delta_{D=0}$ and $\delta_{D=Dn}$ as a Function of f_M for Study $1^{CM\leq C}$. When C_M is not an accurate measure of compliance, the relative performance of **BA**, **CD**, and **BSR** depends on the fraction of subjects with C_M data.

not use C_M data perform well. **ITT** performs the best because it only relies on responses in subjects assigned to D_n=0 to estimate P(Y|D=0). **BSR** performs well for the same reason as **ITT**, plus, C_Q=0 is a valid measure of compliance.

Since C_M=0 is a less accurate measure of compliance than C_M=1, (under $C_{M,M\le C}$, it is equally probably that a fully compliant subject has C_M=0, C_M=.5, or C_M=1 but a noncompliant subject only has C_M=0), the outcome of one third of subjects with D=D_n contributes to **ALL**'s estimate of P(Y|D=0). Thus, the average response of subjects assigned to placebo appears better than the true average outcome at zero dose—contributing to error in estimated drug effect.

6.3.2 Investigation of Performance at Specific Locations in Parameter Space: Self-Reported Compliance is Accurate—Study $2^{CM\le C}$

Given that questionnaires, electronic caps, and pill counts measure different aspects of drug intake, it is possible that their conditions of use determine which tool yields a more accurate measure of compliance. The purpose of this experiment is to investigate performance under a worst-case scenario for methods that assume C_Q is a less accurate measure of compliance than C_M. Here, an extreme case is considered: there is perfect agreement between C_Q and C, but C_M underestimates C. This study is identical to study 2, except that P(C_M|C)=$C_{M,M\le C}$.

Table 6.5 indicates the range of parameters sampled for this study. Note that the boldface value is the only setting differing from study 2 (compare to Table 5.4).

Table 6.5. Distribution of Random Parameter Values in Study $2^{CM\leq C}$.

Parameter	**Distribution**
N	U(50,400)
f_M	U(0,1)
a, *b*	Rather than varying *a* and *b*, $P(C_Q\|C)$ is constructed by directly setting $WT_{Q=C}$, $WT_{Q\perp C}$, and $WT_{Q\geq C}$. $WT_{Q=C}=1$, $WT_{Q\perp C}=0$, and $WT_{Q\geq C}=0$
ρ	U(logit(.1),logit(.5))
$WT_{M\leq C}$	**1**

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 6.6.

Figure 6.3, the boxplot of Δ_R for study $2^{CM\leq C}$, shows how important it is to choose the gold standard correctly. Here, **BA** and **BSR** yield better estimates of exposure-response than a method depending on a more costly experimental design—**ALL**. **ML** has access to the same data as **BSR**, but yields P(Y|D) estimates that are poorer, even, than **ALL**. By considering C_Q as a fallible measure of C_M, **ML**, in a sense, suffers from model misspecification.

Comparison of the plot of $\delta_{D=Dn}$ as a function of f_M in study $2^{CM\leq C}$ to the plot of $\delta_{D=Dn}$ as a function of f_M in study 2 shows that **BSR**'s performance when $C_M\leq C$ (Figure 6.4) is equivalent to **ALL**'s performance when $C_M=C$ (Figure 5.6). The plots also highlight a dramatic change in **ML**'s performance. In Figure 5.6, **ML**'s estimates of

Table 6.6. Observed Range of Parameter Values in Study $2^{CM \le C}$.

Parameter	**Range**
N	51—400
f_M	0—1
cor(C_Q,C)	1
cor(C_M,C)	**.424—.759**
P(Y|D=0)	.101—.499
P(Y|D=D_n)	.501—.899
P(C)	P(C=0): .173—.453, P(C=.5): .227—.558, P(C=1): .182—.446

P(Y|D=D_n) are nearly equivalent to those of **ALL**. In Figure 6.4, **ML**'s estimates of P(Y|D=D_n) do not approach those of **ALL** until f_M>.9. And **ALL** is not the best method of analysis.

Study $2^{CM \le C}$ represents a best-case scenario for **BSR**. However, study $1^{CM \le C}$ may be a more realistic design since it allows for error in both self-reported and electronically monitored compliance. In this case, **ML** performs best. It is interesting to note that **ITT**, a method favored in practice because it is believed to be robust to assumptions, fails to demonstrate any advantage here.

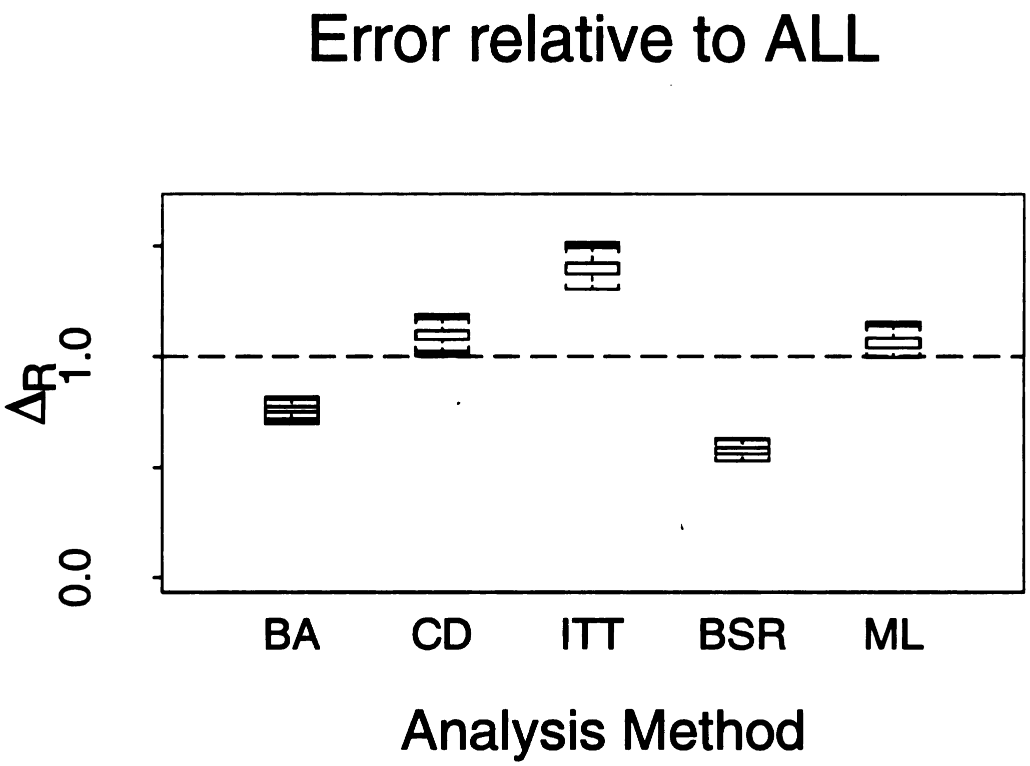


Figure 6.3. Boxplot of Δ_R for Study $2^{CM \le C}$. When C_Q is an accurate measure of compliance but C_M is biased, **BSR** is the best method of analysis. Although **ML** has just as much C_Q data as **BSR**, it performs poorer than **ALL** because it uses C_M as the gold standard.

6.4 Investigation of Sensitivity to *A2* (Random Assignment to M)

An implication of *A2*, the assumption of random assignment to M, is that self-reported compliance is independent of the availability of C_M data. Formally stated, *A2* implies

$$P(C_Q|C,M) = P(C_Q|C).$$

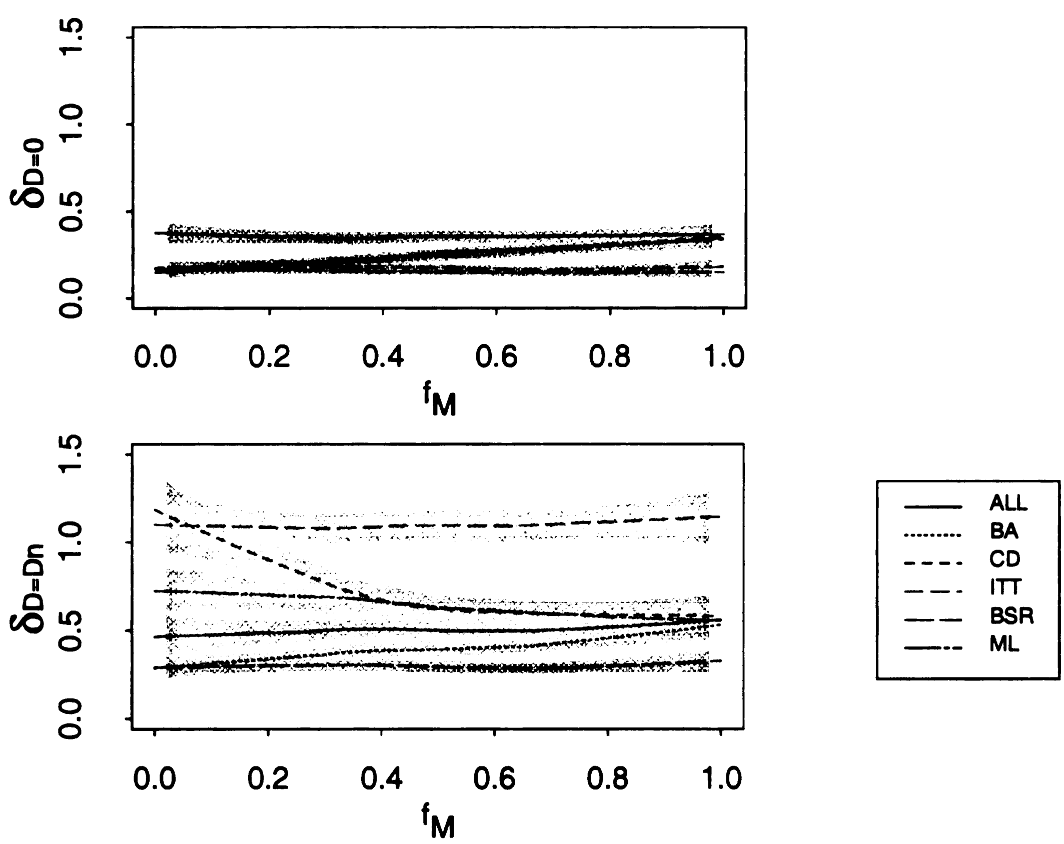


Figure 6.4. Plots of $\delta_{D=0}$ and $\delta_{D=Dn}$ for Study $2^{CM\leq C}$. When C_Q is an accurate measure of compliance but C_M is biased, **ML**'s estimates of P(Y|D=D$_n$) converge on **ALL** for f_M>.9. This is a dramatic change relative to the case in which both C_Q and C_M are accurate—study 2 (Figure 5.6) shows that **ML** is equivalent to **ALL** for nearly all values of f_M.

One may argue that patients who knowingly have compliance monitored by an electronic cap will report their intake more accurately on a questionnaire than patients who do not have a secondary source of information validating their self report. Experimental data are unavailable to determine if this assumption is true. Therefore, sensitivity to *A2* is investigated. To determine robustness to *A2*, C_Q data are simulated from two different models for P(C_Q|C)—the model selected depends on the assigned value of M.

6.4.1 An Extreme Discrepancy in Self-Report Quality for Subjects with M=0 Versus Subjects with M=1; C_Q is Accurate for C in Subjects with M=1, but C_Q has No Information About C in Subjects with M=0—Study $2^{M=0:CQ\perp C,M=1:CQ=C}$

First, an extreme case is considered—self-reported compliance is accurate in subjects who have compliance electronically monitored, but C_Q has no correlation with C in subjects who do not have C_M data. To be more specific, C_Q is simulated from $C_{Q,Q=C}$ for patients with M=1 and C_Q is simulated from $C_{Q,Q\perp C}$ for patients with M=0 (Refer to Table A.1 for the exact probabilities defining $C_{Q,Q=C}$ and $C_{Q,Q\perp C}$).

For the purpose of comparison, parameter values other than those affected by ***A2*** are identical to those in study 2 and study 3. Recall that in study 2, C_Q is perfectly correlated with C in all subjects, not just those with M=1. In study 3, C_Q is has zero correlation with C in all subjects, not just those with M=0. To indicate the relationship with study 2 parameters, this study is referred to as Study $2^{M=0:CQ\perp C,\ M=1:CQ=C}$. The ranges on Study $2^{M=0:CQ\perp C,\ M=1:CQ=C}$ parameter values are listed in Table 6.7. Boldface type is used to indicate which values differ from study 2 (compare to Table 5.4).

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 6.8.

Figure 6.5, a boxplot of Δ_R for study $2^{M=0:CQ\perp C,M=1:CQ=C}$, suggests that **ML** is robust to ***A2***. Although **CD** yields the most precise exposure-response estimates, **ML** has the next lowest Δ_R. A comparison of this plot to the analogous study 2 plot (Figure 5.5) shows that **BA** and **BSR** suffer more from the change to $P(C_Q|C)$ than **ML**. Neither **BA** nor **BSR** calibrates C_Q data; their performance simply reflects the resultant decrease in $cor(C_Q,C)$. Note that **BSR** is equivalent to **BA** because the two methods only differ with respect to the data used to determine exposure in subjects with M=1. Here, C_Q is equivalent to C_M in subjects with M=1.

Table 6.7. Distribution of Random Parameter Values in Study $2^{M=0:CQ\perp C,\ M=1:CQ=C}$.

Parameter	**Distribution**
N	U(50,400)
f_M	U(0,1)
a*, *b	**Rather than varying *a* and *b*, $P(C_Q\|C)$ is constructed by directly setting $WT_{Q=C}$, $WT_{Q\perp C}$, and $WT_{Q\geq C}$ for M=0 and M=1** **M=0: $WT_{Q=C}=0$, $WT_{Q\perp C}=1$, and $WT_{Q\geq C}=0$** **M=1: $WT_{Q=C}=1$, $WT_{Q\perp C}=0$, and $WT_{Q\geq C}=0$**
ρ	U(logit(.1),logit(.5))

Interestingly, a comparison of Figure 6.5 to Figure 5.7 shows that **ML** yields more precise P(Y|D) estimates in study 3—where there is no information in C_Q for any subject—than in study $2^{M=0:CQ\perp C,M=1:CQ=C}$—where C_Q is accurate in subjects with M=1. This demonstrates that **ML** is hurt more by an incorrect calibration between C_Q and C_M, even if there is good information in C_Q and C_M than by a trustworthy calibration that reveals there is no relationship between C_Q and C_M. Although **ML** demonstrates sensitivity to ***A2***, **ML**'s performance is least variable among methods that use C_Q data between the conditions of study 2, study $2^{M=0:CQ\perp C,M=1:CQ=C}$, and study 3.

Table 6.8. Observed Range of Parameter Values in Study $2^{M=0:CQ\perp C,\ M=1:CQ=C}$.

Parameter	**Range**
N	51—400
f_M	0—1
cor(C_Q,C|M=0)	**-.866—.875**
cor(C_Q,C|M=1)	**1**
cor(C_M,C)	1
P(Y|D=0)	.100—.499
P(Y|D=D_n)	.501—.900
P(C)	P(C=0): .173—.472, P(C=.5): .208—.558, P(C=1): .182—.435

6.4.2 A Moderate Discrepancy in Self-Report Quality for Subjects with M=0 and M=1; C_Q is Accurate for C in Subjects with M=1, but C_Q Overestimates C in Subjects with M=0—Study $2^{M=0:\ CQ\geq C,\ M=1:\ CQ=C}$

A less extreme, and, perhaps, more realistic example of a situation in which the validity of C_Q differs between patients with M=0 and M=1 is considered. As in the previous study, C_Q for patients with C_M data is perfectly correlated with C. However, in this example, C_Q is drawn from a distribution in which C_Q is equal to or overestimates C ($C_{Q,Q\geq C}$ in Table A.1) for subjects with M=0. Hence, there is some information on C in C_Q for subjects without C_M data. Aside from the model for P(C_Q|C) in subjects with M=0, all study parameters are as in study 2. The ranges of parameters in this study (Study

$2^{M=0:CQ\geq C,\ M=1:CQ=C}$) are listed in Table 6.9. Boldface type is used to indicate which ranges are different than in study 2 (compare to Table 5.4).

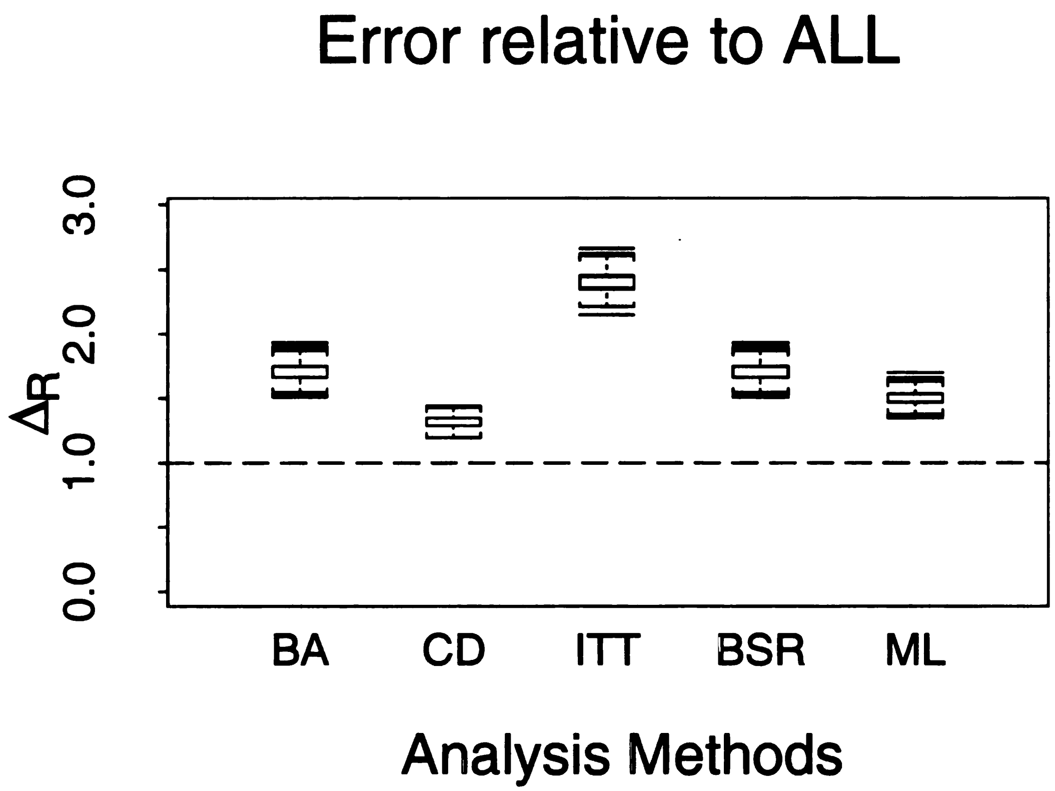


Figure 6.5. Boxplot of Δ_R for Study $2^{M=0:CQ\perp C,M=1:CQ=C}$. ML is robust to a violation of ***A2***. When self-reported compliance is inaccurate in subjects who do not have C_M data, but is accurate in subjects that do, **ML** yields the second best estimates of P(Y|D).

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 6.10.

Table 6.9. Distribution of Random Parameter Values in Study $2^{M=0:CQ\geq C,\ M=1:CQ=C}$.

Parameter	**Distribution**
N	U(50,400)
f_M	U(0,1)
a, *b*	**Rather than varying *a* and *b*, $P(C_Q \mid C)$ is constructed by directly setting $WT_{Q=C}$, $WT_{Q\perp C}$, and $WT_{Q\geq C}$ for M=0 and M=1** **M=0: $WT_{Q=C}=0$, $WT_{Q\perp C}=0$, and $WT_{Q\geq C}=1$** **M=1: $WT_{Q=C}=1$, $WT_{Q\perp C}=0$, and $WT_{Q\geq C}=0$**
ρ	U(logit(.1),logit(.5))

Figure 6.6, the Δ_R plot for study $2^{M=0:CQ\geq C,\ M=1:CQ=C}$, shows that **BA**, **CD**, **BSR**, and **ML** have similar estimation precision in this study. A comparison of Figure 6.6 to the analogous study 2 plot (Figure 5.5) shows that **BA** and **BSR** pay a heavier price than **ML** for the inaccuracy of C_Q in M=0 subjects. A comparison of Figure 6.6 to Figure 6.5 shows that **CD** loses its competitive advantage as the correlation between C_Q and C increases in the M=0 subjects.

6.5 Investigation of Sensitivity to *A2* and *A3* Simultaneously

Study $1^{CM\leq C,\ M=0:CQ\geq C,\ M=1:CQ=C}$ considers a situation in which both *A2* and *A3* are violated simultaneously. That is, the electronic monitor underestimates true intake and self-reported compliance is influenced by the presence of the electronic monitor. The ranges on simulation parameter values are the same as in study 1 except for parameters

Table 6.10. Observed Range of Parameter Values in Study $2^{M=0:CQ\geq C,\ M=1:CQ=C}$.

Parameter	**Range**
N	51—399
f_M	.00291—1
cor(C_Q,C|M=0)	**0—.919**
cor(C_Q,C|M=1)	**1**
cor(C_M,C)	1
P(Y|D=0)	.101—.500
P(Y|D=D_n)	.500—.899
P(C)	P(C=0): .173—.472, P(C=.5): .208—.558, P(C=1): .182—.435

relating to P(C_Q|C) and P(C_M|C). Table 6.11 lists the ranges of parameters used (compare to Table 5.2). The values differing from study 1 are indicated by boldface type.

Observed values of certain parameters in the data actually simulated in a typical study are given in Table 6.12.

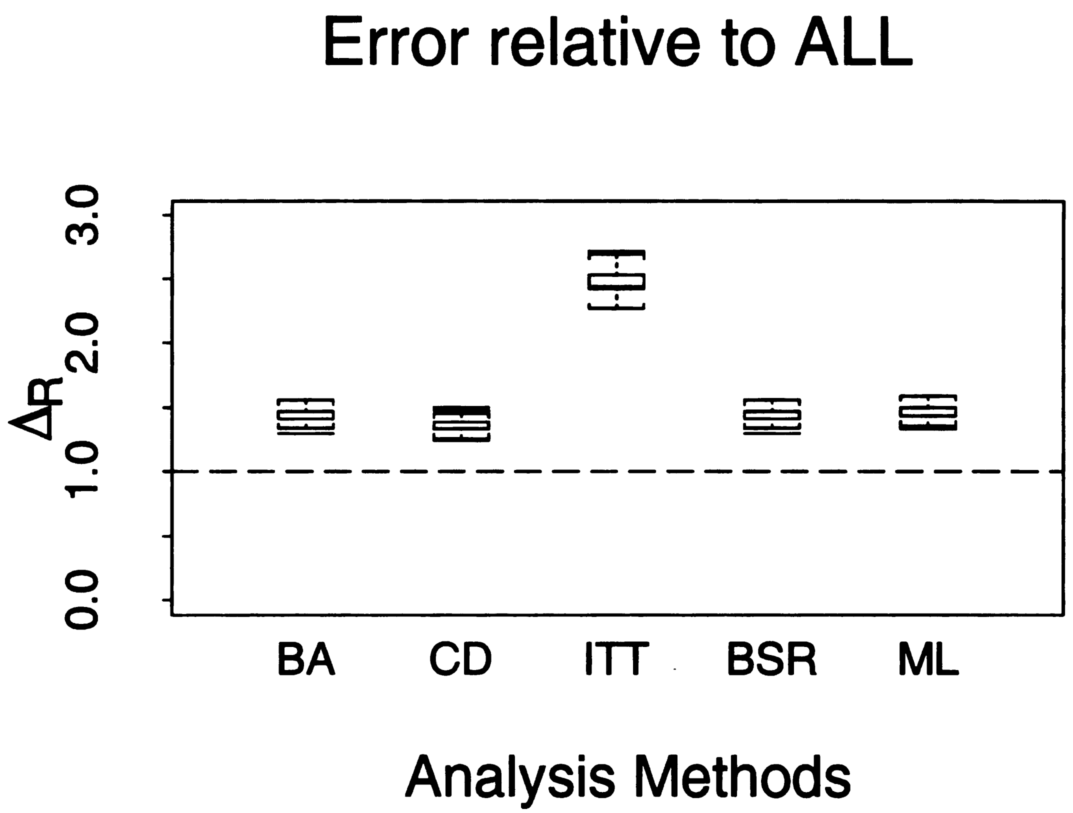


Figure 6.6: Boxplot of Δ_R for Study $2^{M=0:CQ\geq C,\ M=1:CQ=C}$. When C_Q overestimates compliance in subjects with M=0, but accurately measures compliance in subjects that have C_M data, **BA**, **CD**, **BSR**, and **ML** yield equivalent estimates of exposure-response.

Figure 6.7, a boxplot of Δ_R for study $1^{CM\leq C,\ M=0:CQ\geq C,\ M=1:CQ=C}$, reveals that when *A2* and *A3* are simultaneously violated, **BSR** yields the best estimates of exposure-response. A comparison of cor(C_Q,C) and cor(C_M,C) in Table 6.12 reveals why **BSR** performs better than **CD**—on average, C_M is more biased than C_Q. Interestingly, **ML** and **BA** have similar performance. When the correlation between self-reported compliance and true compliance depends on M, **ML** behaves like a method that does not interpret C_Q

data more "intelligently" via calibration. Due to the bias in C_M, the most costly experimental design—**ALL**—is only as good as **BA** and **ML**.

Table 6.11. Distribution of Random Parameter Values in Study $1^{CM\leq C,\ M=0:CQ\geq C,\ M=1:CQ=C}$.

Parameter	**Distribution**
N	U(50,400)
f_M	U(0,1)
$\boldsymbol{a, b}$	**M=0: $WT_{Q=C}=0$, $WT_{Q\perp C}=0$, and $WT_{Q\geq C}=1$**
	M=1: $WT_{Q=C}=1$, $WT_{Q\perp C}=0$, and $WT_{Q\geq C}=0$
ρ	U(logit(.1),logit(.5))
$\mathbf{WT_{M\leq C}}$	**1**

6.6 Sensitivity to Distributional Assumptions—Simulation Studies Drawing from Real Distributions of $P(C_Q|C)$

Up until this point, $P(C_Q|C)$ is selected at random from a wide range of "reasonable" values. Here, sensitivity to the source of $P(C_Q|C)$ is explored. The conditional distribution for C_Q given C is created from a real data set. With C assumed equal to C_M, the model for $P(C_Q|C)$ is constructed by computing the conditional probability matrix given joint observations of C_Q,C_M measured in individuals. Details about the data source are given in Chapter 7.

6.6.1 Source of Data

The exact probabilities of this "real" $P(C_Q|C)$ distribution—referred to as $P(C_Q|C)^{Real}$—is indicated in Table 6.13.

To focus on the influence of the source of $P(C_Q|C)$, all other parameters are generated as in study 1. Table 6.14 lists the parameter values fixed in study $1^{P(CQ|C)Real}$.

Table 6.15 lists the range of parameter values explored in study $1^{P(CQ|C)Real}$. The range of parameter values observed in actual data simulated are listed in Table 6.16.

A comparison of Figure 6.8, the plot of Δ_R for study $1^{P(CQ|C)Real}$, to Figure 5.1, the plot of Δ_R for study 1, shows that the methods yield the same rank order in performance regardless of the source of C_Q and C_M data. **ML** has the greatest precision. **CD** performs nearly as well as **ML**.

Table 6.12. Observed Range of Parameter Values in Study $1^{CM\leq C,\ M=0:CQ\geq C,\ M=1:CQ=C}$.

Parameter	**Range**
N	51—400
f_M	0—1
$cor(C_Q,C)$	.486—1
$cor(C_Q,C\|M=0)$	**0—.919**
$cor(C_Q,C\|M=1)$	**1**
$cor(C_M,C)$	**.424—.733**
$P(Y\|D=0)$	.100—.499
$P(Y\|D=D_n)$	.501—.900
$P(C)$	$P(C=0)$: .173—.472, $P(C=.5)$: .208—.558, $P(C=1)$: .182—.435

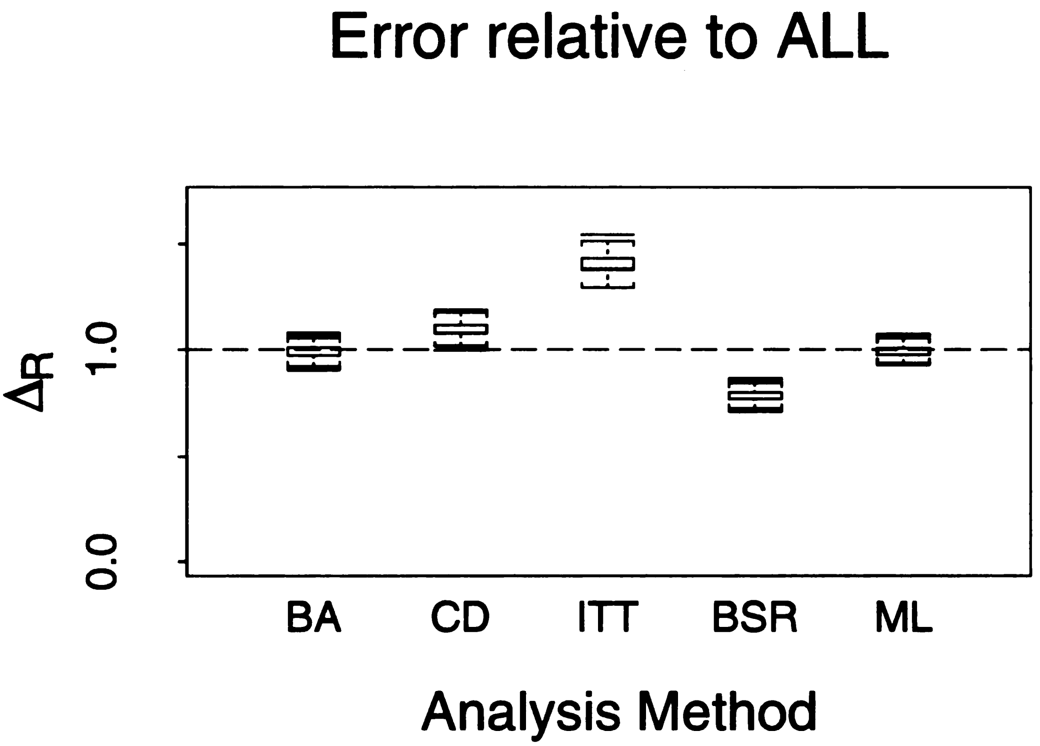


Figure 6.7: Boxplot of Δ_R for Study $1^{M=0:CQ\geq C,\ M=1:CQ=C\ \&\ CM\leq C}$. When ***A2*** and ***A3*** are simultaneously violated, **BSR** yields the most precise estimates of exposure-response. **BA** and **ML** yield estimates of P(Y|D) that are equivalent to **ALL**. **ITT** demonstrates no relative benefit.

Table 6.13. P(C_Q|C) Computed on Real Data.

			C	
		1	2	3
C_Q	1	.385	.04	0
	2	.346	.68	.167
	3	.269	.28	.833

Table 6.14. Fixed Parameter Values in Study $1^{P(C_Q|C)Real}$.

Parameter	**Fixed Value**
N_{lhs}	100 ×5
P(C)	P(C=0)=P(C=.5)=P(C=1) = 1/3
λ_1	.5/3
λ_2	.5/6
θ_{1prior}	$P(Y\|D=0)=P(Y\|D=.5D_n)=P(Y\|D=D_n) = .5$
θ_{2prior}	$P(C=k\|C_Q=j)=1/3$ for j,k {0,.5,1}
$P(C_M\|C)$	$C_{M,M=C}$ (see Figure A.5)

Table 6.15. Distribution of Parameter Values in Study $1^{P(C_Q|C)Real}$.

Parameter	**Range**
N	U(50,400)
f_M	U(0,1)
a, *b*	**NA—C_Q is simulated from $P(C_Q\|C)^{Real}$**
ρ	U(logit(.1), logit(.5))

6.7 Discussion

In this chapter, assumptions about compliance measurement are challenged to determine their influence on method performance. Here, simulated data either violate an assumption unique to **ML** ($P(C_Q|C,M)=P(C_Q|C)$), an assumption of all methods that use

Table 6.16. Observed Range of Parameter Values in Study $1^{P(CQ|C)Real}$.

Parameter	**Range**
N	50—399
f_M	0—1
cor(C_Q,C)	**.243—.791**
P(Y|D=0)	.100—.500
P(Y|D=D_n)	.500—.900
P(C)	P(C=0)= .22—.547, P(C=.5)= .19—.5, P(C=1)= .169—.444

C_M data (C_M=C), or both simultaneously. The assumptions are violated to differing degrees. Extreme violations are used to illustrate worst-case scenarios. More moderate violations suggest more realistic behavior. The two approaches give us a glimpse of the overall variability in method performance.

Study $1^{CM\leq C}$ shows that when both C_M and C_Q measure C with error, **ML** yields estimates of exposure-response that are equivalent to those obtained through a more costly trial design (**ALL**). However, study $2^{CM\leq C}$ warns that if C_M measures C with error but C_Q is a perfect measure of compliance, **BSR** and **BA** are the best methods of analysis. Rather than putting all of the results reported so far in question, study $2^{CM\leq C}$ illustrates how important it is to choose the gold standard correctly. Since **BSR** treats self-reported compliance as a perfect measure of compliance, when C_Q=C, it becomes analogous to **ALL** in studies with C_M=C. **ML**'s performance suffers when the gold standard is chosen

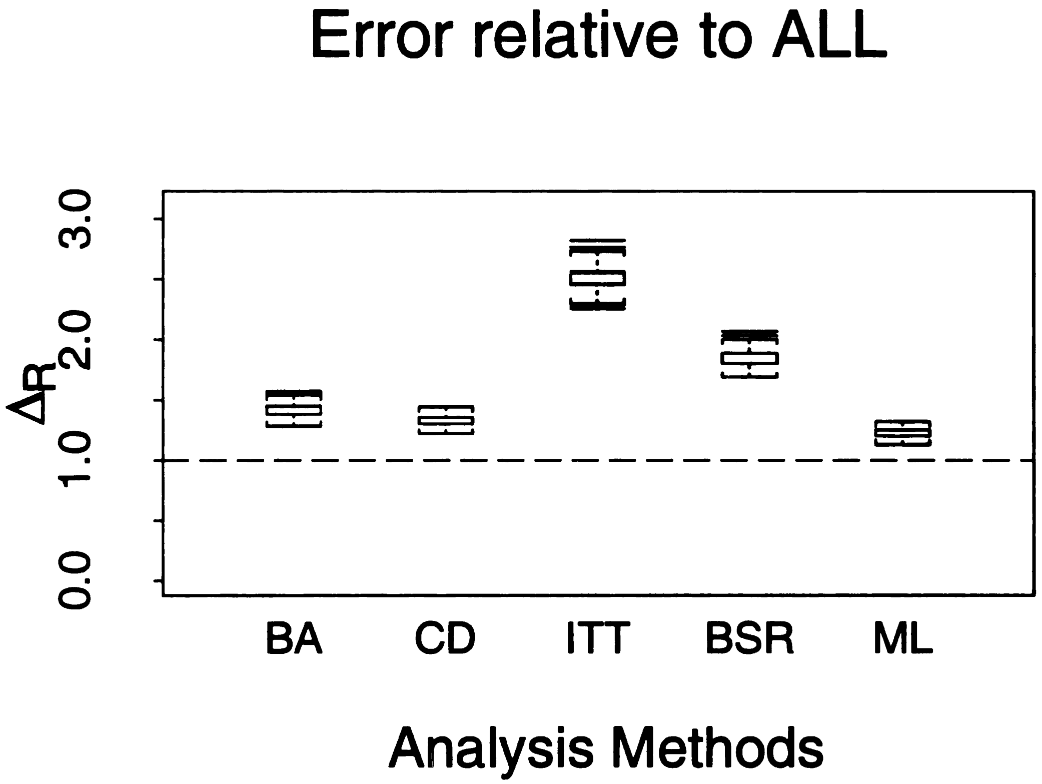


Figure 6.8. Sensitivity to the Source of P(C_Q|C): Study $1^{P(CQ|C)Real}$. When C_Q and C_M are simulated from a model based on published data, relative method performance is the same as when P(C_Q|C) is chosen from a wide range of values (compare to Figure 5.1).

incorrectly, but this does not indicate a weakness in the likelihood it optimizes. Clearly, it is up to the data analyst not to be prejudiced by the terms "self-report" and "electronic monitoring". Expert opinion about which is more accurate should be gathered before commencing data analysis. This is especially important when using pill count data—the difference in accuracy between unannounced pill counts and assessments during a visit to the care provider may be large.

In contrast, ***A2*** presents a challenge to **ML**. Study $2^{M=0:CQ\perp C,M=1:CQ=C}$ reveals that when cor(C_Q,C)=0 in subjects who have no C_M data, but cor(C_Q,C)=1 in subjects who have C_M data, **ML** trails **CD** as the best method of analysis. However, this is an extreme example of a disagreement between the accuracy of C_Q in subjects with M=0 and M=1. In study $2^{M=0:CQ\geq C,\, M=1:CQ=C}$, where cor($C_Q$,C)<1 in subjects who have no C_M data but cor(C_Q,C)=1 in subjects who have C_M data, **ML** is among the best methods of analysis. In study $1^{M=0:CQ\geq C,\, M=1:CQ=C\ \&\ CM\leq C}$, where there is a simultaneous violation of ***A2*** and ***A3***, **ML** trails **BSR** as the best method of analysis.

Although **BSR** and **CD** are each the best method of analysis under different violations of ***A2*** and ***A3***, **ML** is the next best method of analysis in both cases. Since the extent to which ***A2*** and ***A3*** are incorrect is unknown, **ML** seems to be the safest bet.

It is unknown whether ***A2*** is of genuine concern. The validity of ***A2*** likely depends on the cause of inaccuracy in C_Q. If error in self-reported compliance is primarily due to patient lying, then C_Q is something patients can manipulate. That is, the presence of an electronic monitor can influence C_Q. However, if C_Q is an inexact measure of C because patients forget how many pills they do not remember to take (as many believe(Chesney 2000)), an electronic recording device hidden in the cap of a pill bottle is not likely to change this. After all, if it had an important impact on memory, the manufacturers of the device would have a much more valuable commodity in their possession! Of course, the presence of the device may lead patients to meticulously keep a diary of their intake. However, this is likely to be exceptional, not normal, behavior.

In contrast, ***A3*** is likely untrue. The likelihood (i.e. model for the data) can, however, be adjusted to allow for error in C_M. Since **ML** performs quite well without such a change, its performance can only improve as the model better reflects the realities of the relationship between C and C_M.

Study $1^{P(CQ|C)Real}$ shows that **ML** performs well when self-reported compliance data are drawn from a more realistic source of $P(C_Q|C)$. However, the most realistic evaluation may be to compare method performance when C_Q,C_M, and Y are drawn from a real data set.

6.8 References

Angrist, J. D., G. W. Imbens and D. R. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables." Journal of the American Statistical Association **91**: 444-472.

Bangsberg, D. R., F. M. Hecht, E. D. Charlebois, A. R. Zolopa, M. Holodniy, L. B. Sheiner, J. D. Bamberger, M. A. Chesney and A. Moss (2000). "Adherence to Protease Inhibitors, HIV-1 Viral Load, and Development of Drug Resistance in an Indigent Population." AIDS **14**(4): 357-366.

Chesney, M. A. (2000). "Factors Affecting Adherence to Antiretroviral Therapy." Clinical Infectious Diseases. Supplement. **30**: S171-S176.

Efron, B. and D. Feldman (1991). "Compliance as an Explanatory Variable in Clinical Trials." Journal of the American Statistical Association **86**: 9-22.

Goetghebeur, E., G. Molenberghs and J. Katz (1998). "Causal Effect of Compliance on Binary Outcome in Randomized Controlled Trials." Statistics in Medicine **17**: 341-355.

Robins, J. M. (1998). "Correction for Non-Compliance in Equivalence Trials." Statistics in Medicine **17**: 269-302.

Sheiner, L. B. and D. B. Rubin (1995). "Intention-to-Treat Analysis and the Goals of Clinical Trials." Clinical Pharmacology and Therapeutics **57**(1): 6-15.

Turner, B. J. and F. M. Hecht (2001). "Improving on a Coin Toss to Predict Patient Adherence to Medications." Annals of Internal Medicine **134**(10): 1004-1006.

Chapter 7: Analysis of Clinical Data

Abstract

In this chapter, the data analysis methods are applied to a real data set. The data were collected during a clinical trial investigating compliance with and response to protease inhibitors in 34 subjects. The percent of prescribed doses taken was quantified using three compliance-measuring tools in all subjects—a questionnaire, an electronically monitored cap, and an unnanounced pill count. The results of this chapter show that method performance determined using real data is consistent with performance using simulated data.

7.1 Purpose

Thus far, method performance has been evaluated using artificial data. In chapters 5 and 6, **BA**, **CD**, **ITT**, **BSR**, and **ML** are compared with respect to their ability to analyze data generated from P(C), P(C_M|C), P(C_Q|C), and P(Y|D) distributions deemed "reasonable". Simulation studies may illuminate the salient features of method performance, but, nonetheless, are suspect. After all, real data sets may have complex interactions among parameters. These interactions may be intentionally left out of data simulation in order to focus on the factors of interest or their influence may be unknown to the investigator. Another necessary investigation of a proposed analysis method, then, is to determine how well it handles real data. The goal of the investigation presented in this chapter is to determine if the relative method performance evaluated on simulated data extends to a clinical analysis.

7.2 Data Source

Individuals infected with the Human Immunodeficiency Virus (HIV) take approximately two dozen tablets per day(Chesney, Morin et al. 2000). The deadly virus must be treated with a combination of therapies, many of which have short half lives. Further complicating the picture, patients may also take drugs to treat opportunistic infections and conditions unrelated to HIV infection. Since patient compliance is known to decrease with increasing regimen complexity, compliance is important to monitor in the clinical management of patients infected with HIV.

Bangsberg and colleagues measured the percent of prescribed protease inhibitor doses taken in 34 HIV infected homeless people(Bangsberg, Hecht et al. 2000). Compliance was assessed via questionnaire, electronically monitored pill caps, and unannounced pill counts. Subjects were observed for a median of 66 days and were scheduled to have three compliance assessments—with all three tools each time—across the study's duration.

The questionnaire required subjects to self-report how many of the prescribed doses were missed over the previous three days. Bangsberg and coworkers (2000) transformed the self-reported compliance data into a fractional compliance by dividing the number of tablets that should have been ingested during the previous three days less the number that were reported to have been missed by the number of pills that should have been ingested. Electronically monitored compliance data were transformed into a fractional compliance by dividing the actual number of pill bottle openings by the nominal number of openings. The statistic is referred to as the AEMD, or the Adjusted Electronically Monitored Dosing, since Bangsberg and coworkers use self-reported compliance to correct electronically monitored values to reflect proper intake if subjects admit taking doses from a source other than the electronically monitored container. For instance, if a patient reports removing all doses for the entire day at one opening, rather than dispensing from the monitored bottle each time, then the electronic record is adjusted to reflect good compliance. Pill count compliance was computed by dividing the number of doses that disappeared from the subject's possession between two assessments by the number of doses that should have disappeared if the subject had been perfectly compliant with the prescribed regimen.

Bangsberg and coworkers' percent compliance data, as measured by self-report, electronic monitor, and pill count, are used here to evaluate method performance. To be consistent with the notation in chapters 5 and 6, "C_Q" and "C_M" symbolize self-reported and electronically monitored compliance, respectively. Pill-count compliance is represented by "C_{PC}". Here, patient values of C_{PC} are treated as the response (Y) to be predicted using C_M and C_Q data.

As required by the analysis methods, the data are transformed from the reported 0-100% scale to a categorical value. Electronically monitored compliance is categorized by creating a discrete distribution with equal probability of having each C_M value. The cutoff values for C_M in the data that yield such a distribution are as follows.

Compliance category (C_M)	Percent of prescribed drug taken
1	(,50]
2	(50—90]
3	(90,)

The use of parentheses indicates an open interval, while the bracket indicates a closed interval. The same cutoffs are applied in categorizing C_Q values. If C_Q is equivalent to C_M, $P(C_Q)$ will assign subjects to each possible discrete value of C_Q with a 1/3 probability, as well. However, $P(C_Q)$ does not have the same distribution as $P(C_M)$—on average, it overestimates C_M. C_{PC} is transformed into a binary variable—comply ("success") or not comply ("failure")—such that $P(C_{PC})$ is a uniform distribution. The cutoff value of 80%, indicated below, yields a data set in which half of the subjects have Y=0 and half have Y=1.

Response (Y)	C_{PC}
0	(,80]
1	(80,)

Note that the correlation between C_Q and C_M in the continuous data is .60 and the correlation between C_Q and C_M in the categorically transformed compliance data is .54.

7.3 Methods

Here, as in chapters 5 and 6, the goal is to estimate $P(Y|D)$. Except, now, C_M, C_Q, and Y are not simulated. They are taken from a data set of joint (C_M,C_Q,C_{PC}) values. An empirical distribution for $P(C_M,C_Q,C_{PC})$ is assembled from Bangsberg and coworkers' data. Of 102 possible records—based on 34 subjects observed three times—75 records consisting of complete C_M,C_Q,C_{PC} data are culled from the data set. Note that a given individual in the study may contribute zero, one, two, or three (C_M,C_Q,C_{PC}) records to $P(C_M,C_Q,C_{PC})$. For the purposes of the present investigation, each compliance monitoring event is treated as if it arises from a different individual.

First, the 75 data records in P(C_M,C_Q,C_{PC}) are sampled with replacement to yield a new data set with N=75. Next, N_M subjects of the N subjects are randomly selected to have C_M observed. This data set is now exactly like data sets analyzed in chapters 5 and 6 and is analyzed by all methods of analysis.

This procedure is repeated one hundred times for each setting of N_M at 15, 25, 50, and 60.

7.4 Results

Since the data are real, the true P(Y|D) is unknown and its estimation error cannot be calculated. Therefore, a histogram of each method's estimates of P(Y|D=D_n) is presented. Since electronic monitors assess compliance more objectively than self-report, C_M is treated as a better estimate of C. Therefore, **ALL** is considered the gold standard for comparing method performance. Vertical bars demarcating the 5th and 95th percentile of **ALL**'s estimates are shown and the fraction of each method's estimates falling outside of this region is reported. Four sets of histograms are presented—one for each value of N_M investigated.

Figure 7.1 shows that **ML** is essentially equivalent to **ALL** when only twenty percent of subjects have C_M data. That is, .11 of **ML**'s exposure-response estimates fall in the region where ten percent of **ALL**'s estimates lie. The next best method of analysis—**CD**—has thirteen percent more P(Y|D) estimates exceeding **ALL**'s "90th percentile confidence" region. **BA**, **BSR**, and **ITT** have little to no overlap with **ALL**. The cost of **ALL** is only justifiable if methods other than **ML** are to be used.

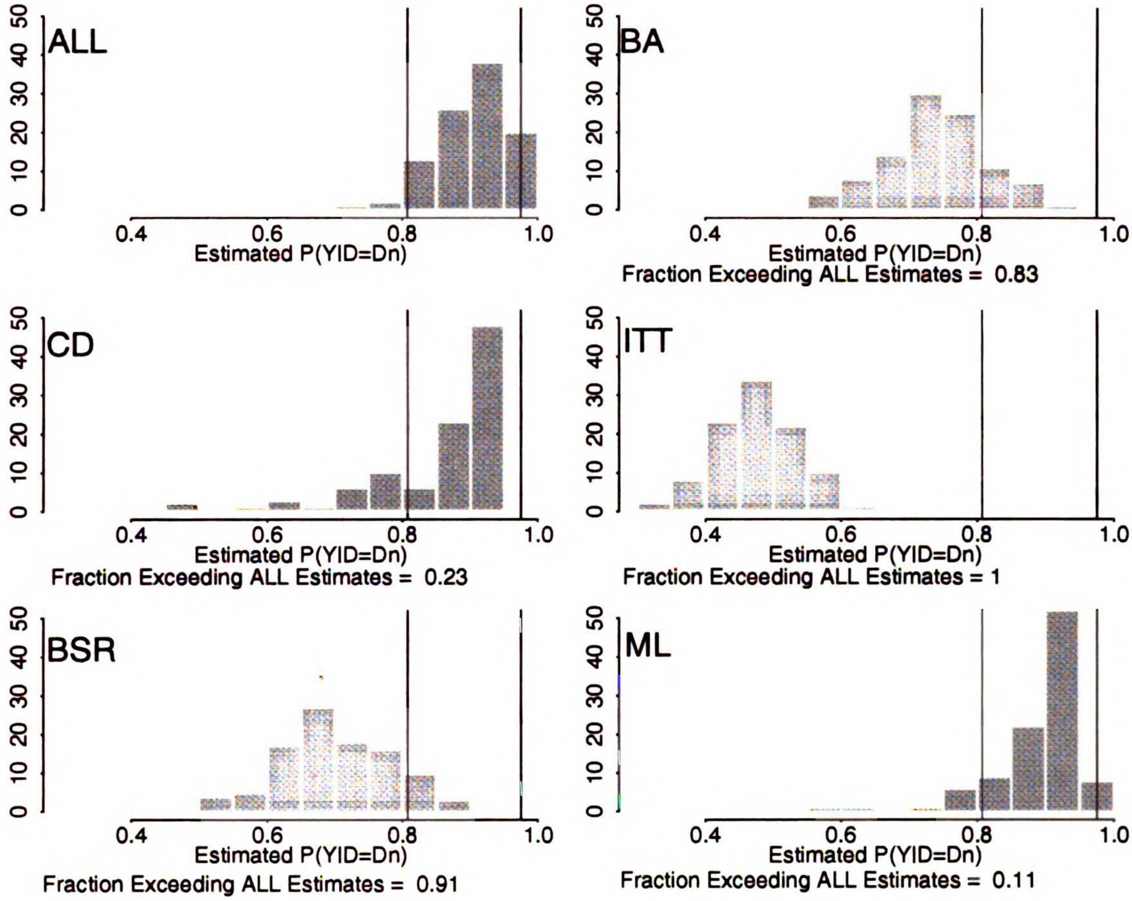


Figure 7.1. Histograms of $P(Y|D=D_n)$ estimates when $N_M=15$ ($f_M=.2$). The high cost of measuring C_M in all subjects is not justified if **ML** is used. **ML** only fails to estimate exposure-response within **ALL**'s 90th percentile confidence region for 1 clinical trial out of 100. By throwing away data in subjects with M=0, **CD** fails to estimate P(Y|D) within **ALL**'s 90th percentile confidence region 13 percent of the time. **ITT** never yields estimates of exposure-response within **ALL**'s 90th percentile confidence region.

7.5 Discussion

Figures 7.2, 7.3, and 7.4 show that method behavior is consistent with the studies on simulated data. **BA** and **CD** improve as f_M increases. **BSR** and **ITT** yield considerably different estimates of exposure-response than **ALL**. **ML** is the best method of analysis. If the next best method, **CD**, is to be used, one needs f_M=2/3 to yield estimates of P(Y|D) that are as good as those returned by **ML** with only f_M=.2.

Most surprisingly, **ML** performs better in these studies relative to all other methods than in studies using simulated data. Here, there is little distinction between exposure-response estimates by **ML** and **ALL**, but there is a great distinction between P(Y|D) estimates for **ML** versus **BA**, **CD**, **ITT**, and **BSR**. While this result should not be interpreted as evidence that **ML** is equivalent to **ALL**, it does suggest that the simulation studies may have been more conservative for **ML** than the studies with real data. Thus, **ML** has not been given too much of an unfair advantage. The lack of a difference in **ML**'s performance using real compliance data makes one more confident in the assumption that error in C_Q and C_M is uncorrelated (among other assumptions).

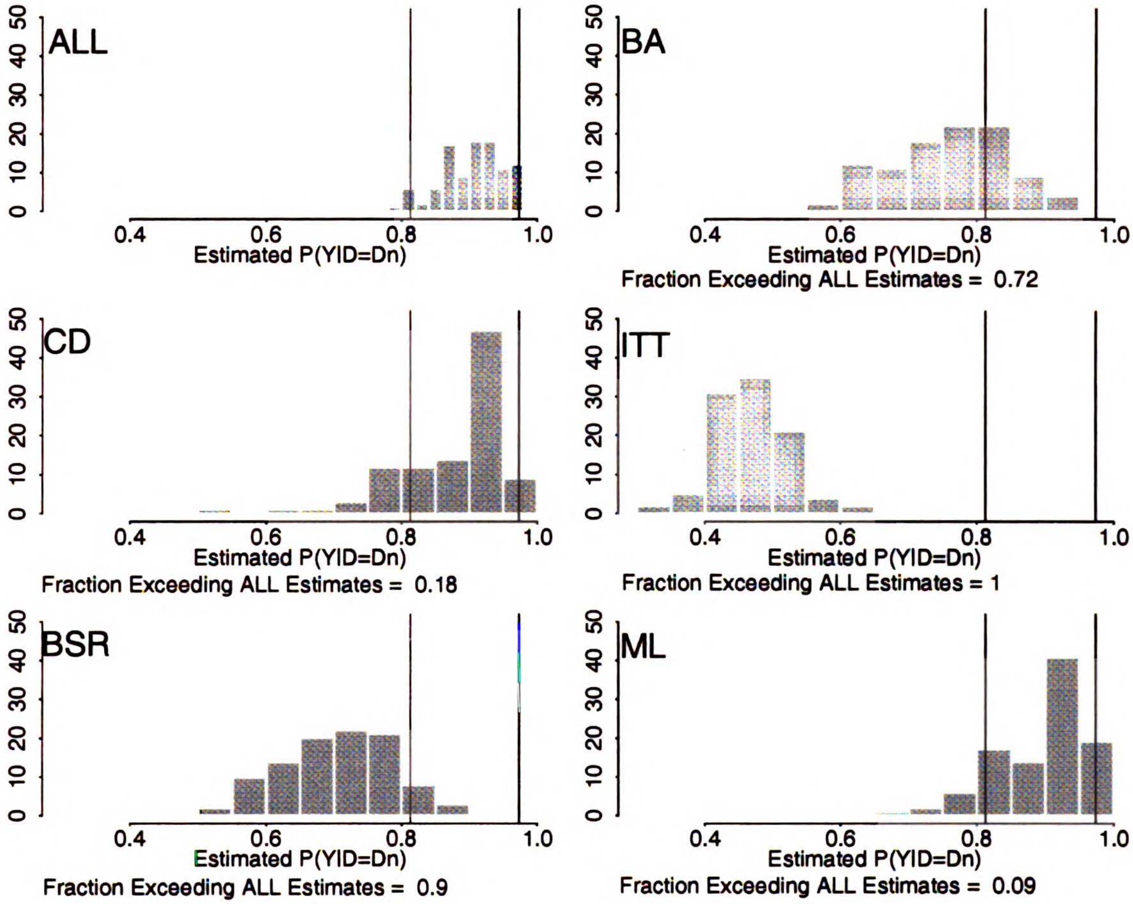


Figure 7.2. Histograms of $P(Y|D=D_n)$ estimates when N_M=25 (f_M=1/3). By throwing away data in the subjects with M=0, C_M fails to estimate P(YID) within **ALL**'s 90th percentile confidence region 8 percent of the time. **ML** is equivalent to **ALL**.

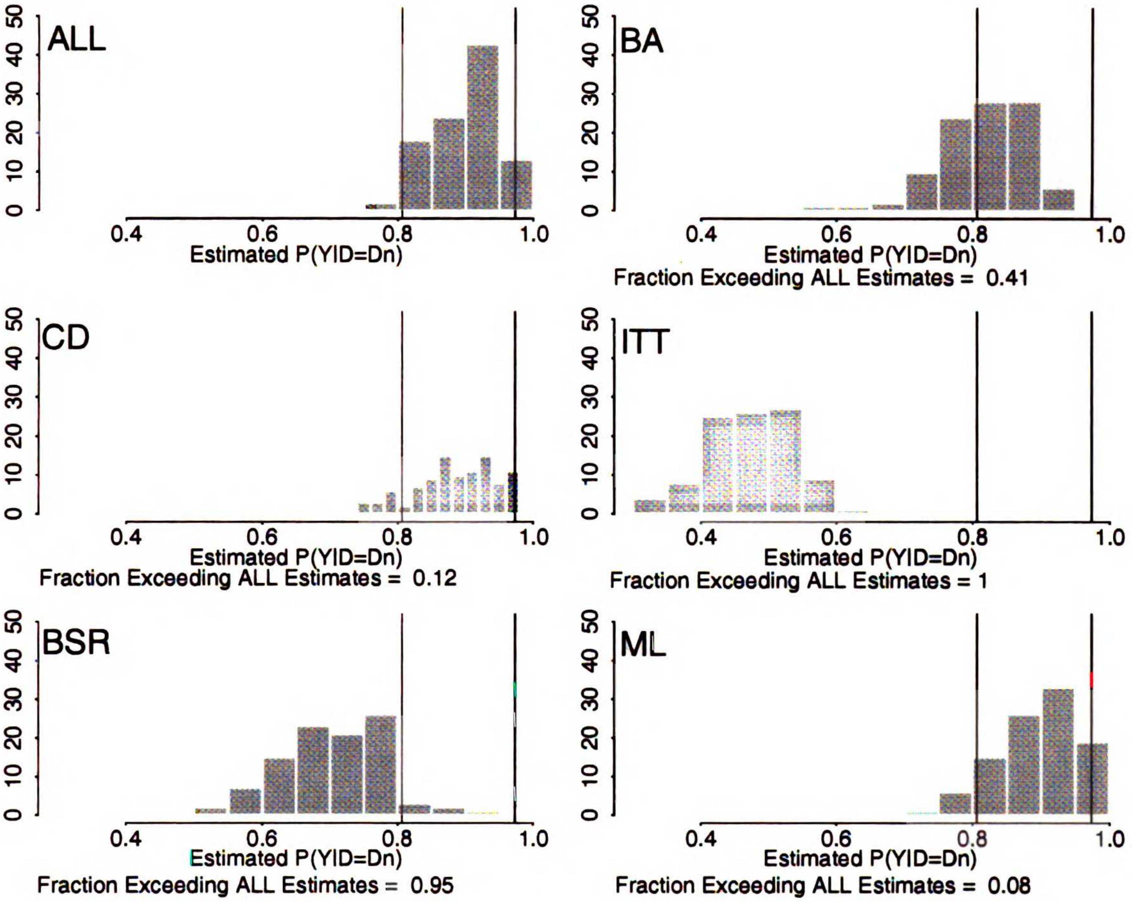


Figure 7.3. Histograms of P(Y|D=D_n) estimates when N_M=50 (f_M=2/3). Despite throwing away data in 1/3 of subjects, **CD** is equivalent to **ALL**.

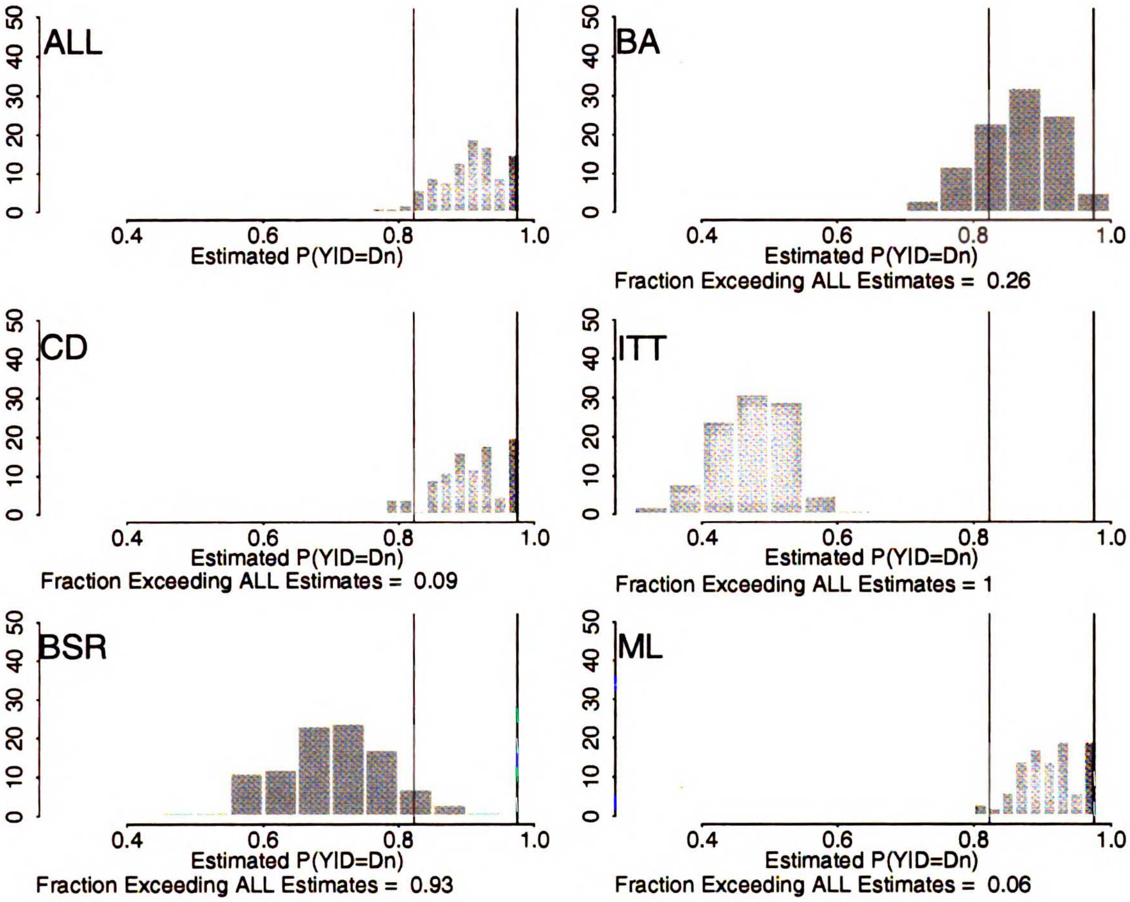


Figure 7.4. Histograms of $P(Y|D=D_n)$ estimates when N_M=60 (f_M=.8). Even with C_M data available in 80% of subjects, **BA** fails to yield estimates of exposure-response within **ALL**'s 90th percentile confidence region for 16 percent of clinical trials.

7.6 References

Bangsberg, D. R., F. M. Hecht, E. D. Charlebois, A. R. Zolopa, M. Holodniy, L. B. Sheiner, J. D. Bamberger, M. A. Chesney and A. Moss (2000). "Adherence to Protease Inhibitors, HIV-1 Viral Load, and Development of Drug Resistance in an Indigent Population." AIDS **14**(4): 357-366.

Chesney, M. A., M. Morin and L. Sherr (2000). "Adherence to HIV Combination Therapy." Social Science and Medicine **50**: 1599-1605.

SECTION III

CONCLUSIONS AND PERSPECTIVES

Chapter 8: General Findings and Recommendations

Abstract

The methods and results presented in this thesis are critically evaluated in this chapter. The goal is to determine whether the results accurately represent method performance within the parameter space explored and, if they do, ascertain the generality of conclusions reached. Here it is argued that even for data types not investigated in this report, the efficiency gained by calibrating biased compliance information to accurate compliance data (under a double sampling scheme) outweighs the consequent risk of obtaining poorer estimates of exposure-response at extreme regions of the parameter space. With the exception of trivial cases (e.g. f_M assumes the value of 0 or 1), however, the results are not diagnostic of the exact conditions under which one method of analysis outperforms another if the data differ in format from those considered here.

8.1 Introduction

As with most scientific experiments, it is necessary to be reductionist when designing computer simulation studies. After all, one has a finite amount of time to spend performing computations. This is not necessarily a limitation. Uncomplicated systems may promote valuable conceptual understanding. The key is to preserve the interesting aspects of a problem while paring down the universe of possible models to satisfy practical constraints.

The goal of this project, as stated in Chapter 3, is to determine the operating characteristics of various methods for estimating exposure-response given a clinical trial in which exposure (compliance) is measured using a biased tool in all subjects and an accurate tool in a random subset. Thus, the format of compliance data is the priority. Saturated models for patient responses are used to generate data under a simple clinical trial design having accurate compliance information missing in a subset. This system is presented in Chapter 4 and in the Appendix.

Specific investigations of the system, presented in Chapters 5 and 6, reveal that among all methods compared, on average, **ML** yields estimates of exposure-response with the least error, has the most favorable power to reject the null hypothesis, and is robust to violations of assumptions about compliance. Chapter 7 shows that **ML** analysis of a real data set is consistent with the results based on simulated data. The work suggests that if one is handed a data set like those analyzed in Chapters 5-7, **ML** is the most efficient method to use.

But exposure and response data in clinical trials are often continuous variables likely to arise from a more complex system than represented by the saturated models used in this investigation. Therefore, an important question about method performance remains to be answered—what is the impact of this work on the analysis of clinical trials with designs differing from those investigated in this report?

8.2 Critique of the Simulation Study Design

In simulation studies, the parameter space explored and the presentation of results can bias method performance. Therefore, it is important to consider whether any of these choices unfairly favor **ML**.

8.2.1 Parameter Space Explored

Random sampling of parameter space takes some control over simulations away from the investigator, thus, reducing bias. A method of stratified sampling, Latin Hypercube Sampling (LHS) yields unbiased estimates of the mean more efficiently than simple random sampling (SRS), so LHS is used in this investigation(McKay, Beckman et al. 1979). By guaranteeing that points are spread out in multidimensional simulation space, LHS reduces the possibility that parameter values cluster in a region that benefits one method over another.

Estimation precision is influenced by the number of simulations performed—the ability to discriminate between methods improves as the number of sets of parameter values explored increases. One may choose the number of parameter sets to obtain estimates with a particular level of precision. Here, the number of iterations is dictated by the amount of time it takes to run each clinical trial simulation and data analysis. In this investigation, 500 sets of parameter values are chosen by LHS. If anything, the number of simulations carried out hurts **ML**'s performance. In Figure 5.3 and Figure 5.6 **ML** and **CD** yield similar estimates of $P(Y|D=D_n)$.

LHS guarantees that points are spread out in multidimensional parameter space, but it does not assure that clinically relevant parameter space is searched, or if it is, that appropriate "weight" is given to its sub-regions. These characteristics are under the control of the investigator. Therefore, it is important to evaluate the influence of the region explored on method performance.

It is difficult to find a typical value of f_M used in practice because there are no available guidelines for computing it. A survey of AIDS Clinical Trials Group (ACTG) designs reveals wide variation in this parameter. In this investigation, f_M spans the range of all possible values (0 to 1), so there is no concern that the range is chosen to favor one method over another.

The study size ranges from 50-400—a span chosen to reflect N in ACTG studies and smaller phase 3 studies. The values of f_M and, perhaps, N represent nonasymptotic conditions. It is known that **ML** performs poorly under nonasymptotic conditions(Little and Rubin 1987). Therefore, the choice of N and f_M likely hurts **ML** (and **CD**).

An infinite selection of possible models for $P(C_Q|C)$ necessitates focusing the investigation on those that are clinically relevant. Since true compliance is unmeasurable, some assumptions are made about what constitutes clinically relevant. Since a low C_Q value is likely more valid than a high C_Q value, a model for $P(C_Q|C)$ in which self-reported intake is equivalent to or an overestimate of compliance is reasonable for simulating C_Q given C. Rather than confining the investigation to just a few such models, a wide range of possible models are created by averaging together two extreme possibilities for $P(C_Q|C)$ (C_Q=C and $C_Q \perp C$) with a model for $P(C_Q|C)$ in which $C_Q \geq C$. This approach seems reasonable given the value of $P(C_Q|C)$ computed on real data presented in Table 6.13.

Figure 8.1 shows the distribution of $cor(C_Q,C)$ values computed on study 1 data. (Note that Figure 8.1 is representative of $cor(C_Q,C)$ in all studies for which $P(C_Q|C)$ is a free parameter.) The vertical bar on the plot at $cor(C_Q,C)=.55$ indicates the $cor(C_Q,C_M)$ observed in pooled literature data sets(Burney, Krishnan et al. 1996; Straka, Fish et al. 1997). This value is similar to $cor(C_Q,C_M)$ computed on the data analyzed in Chapter 7. Assuming that $cor(C_Q,C_M)$ is an overestimate of $cor(C_Q,C)$, it is reasonable that the bulk of the distribution of cor(C_Q,C) values fall below this line. Since **ML**'s performance improves as cor(C_Q,C) increases, **ML** is not helped by such midrange, albeit, clinically relevant, values of cor(C_Q,C).

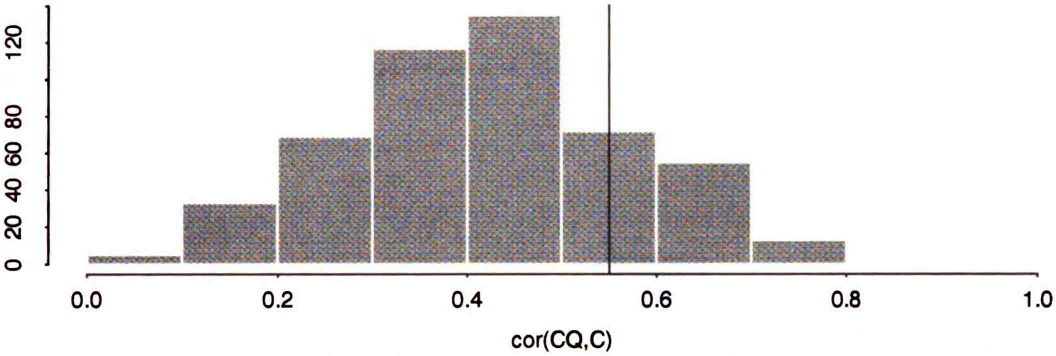


Figure 8.1. Histogram of cor(C_Q,C) in Study 1 Data. The correlation between self-reported compliance and true intake is not uniformly distributed between 0 and .8 in study 1. Half of all simulations have cor(C_Q,C) ranging from .3-.5. The vertical bar at cor(C_Q,C)=.55 represents the correlation between self-reported compliance in data from the literature. This value of cor(C_Q,C_M) is similar to that observed in the data in Chapter 7 and is considered an overestimate of cor(C_Q,C).

As with P(C_Q|C), the model for P(C_M|C) is restricted to clinically relevant values during sensitivity analysis. C_M is thought to underestimate compliance, but have a greater correlation with C than cor(C_Q,C)(Burney, Krishnan et al. 1996; Bangsberg, Hecht et al. 2000; Turner and Hecht 2001). Subjects with perfect compliance are simulated with all possible C_M values, but subjects with zero intake can only have C_M=0. One can be perfectly compliant and simply neglect to use the electronically monitored bottle to dispense tablets, but it is assumed that subjects who don't take drug don't open and close the medication container at all.

Figure 8.2 shows the distribution of cor(C_M,C) values in study $2^{CM \leq C}$ data. The mode of the distribution is skewed slightly higher than the correlation between C_Q and C_M in ACTG data. This is a clinically reasonable assumption. **BA**, **CD**, and **ML** are sensitive to cor(C_M,C), thus **ML** is not favored by the choice of P(C_M|C).

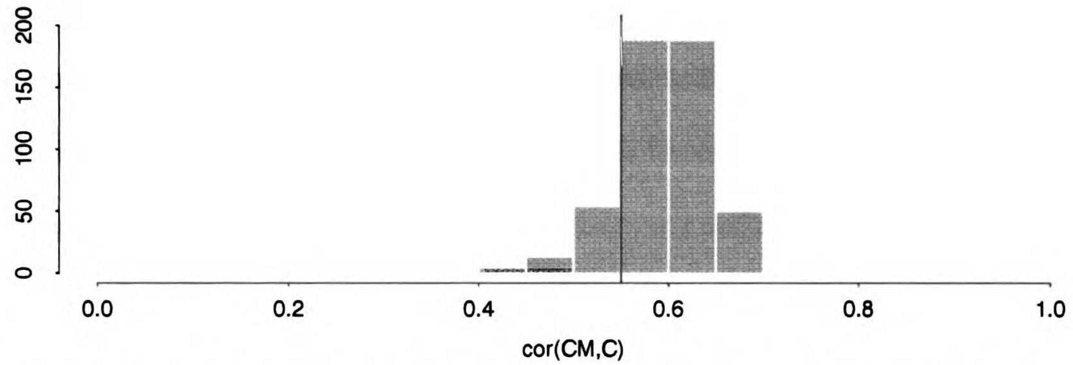


Figure 8.2. Histogram of cor(C_M,C) in Study $2^{CM \leq C}$ Data. The correlation between electronically monitored compliance and true compliance is not uniformly distributed between .4 and .7 in the investigation of sensitivity to C_M=C. The mode of the distribution is skewed slightly higher than cor(C_M,C)=.55—the cor(C_M,C_Q) in the data sets taken from the literature and a value thought to underestimate cor(C_M,C).

Note that in contrast to the method for generating P(C_Q|C) at random, only one model for P(C_M|C) is used. This explains why the variance in cor(C_M,C) is less than the variance in cor(C_Q,C). The difference in variance is reasonable given that C_Q is measured using less standardized tools than C_M. The quality of information in questionnaires and diaries is expected to vary depending on the clinical trial design and the relationship between the subject and the investigator collecting the information(Kaplan and Simon 1990; Catania, Binson et al. 1996; Ickovic and Meisler 1997).

As explained in the Appendix, the model for P(Y|D) is log linear and P(Y|D=.5D_n) is fixed to .5. Randomly drawn values of log(P(Y=1|D=0)) specify the relationship between Y and D. Since **ML** performs best relative to other methods when drug effect is high, **ML** can be made to perform better by having a great number of studies with P(Y=1|D=0) skewed toward .1. Figure 8.3, a histogram of P(Y=1|D=0) of simulation in study 1, shows that the values are nearly uniformly distributed. The slight skewness towards a large drug effect is not unreasonable as we are generally interested in evaluating methods with respect to their ability to discover a real drug effect. (After all, all methods are equivalent in their ability to discover a null drug effect.) Therefore, **ML** is not unfairly favored by the selection of P(Y|D).

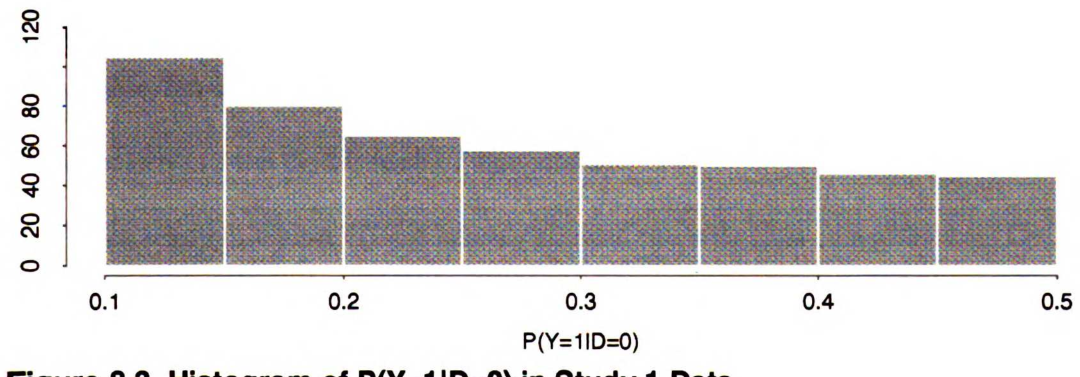


Figure 8.3. Histogram of P(Y=1ID=0) in Study 1 Data.

The study design models, $P(D_n)$ and P(M), are not conditional on any outcome variables. All subjects are equally distributed among placebo and nominal dose groups and subjects with C_M data are equally divided between the D=0 and $D=D_n$ groups. This decision reflects a reasonable clinical trial design. Note that the choice of $P(D_n)$ favors **ITT** by guaranteeing that it has the same number of subjects contributing to the $P(Y|D=0)$ and $P(Y|D=D_n)$ estimates. **ITT** is the only method which enjoys this benefit.

The most questionable parameter value is the arbitrarily chosen P(C). Figure 8.4 shows a histogram of P(C) values observed in study 1 data. In simulation, it is equally probable that a subject has C=0, C=.5, or C=1. Clinically, about one-sixth of patients are poor compliers, another one-sixth are perfect compliers, while the behavior of a full two-thirds falls somewhere in between these extremes(Urquhart 1997). This is determined with respect to the number of drug holidays taken per month as measured via an electronic monitor. Note that since P(C=0) and P(C=1) equivalently overestimate $P(C_M=0)$ and $P(C_M=1)$, the effect on P(C) likely balances out.

Because C_M is believed to underestimate C, P(C) may actually be skewed such that P(C=0)<1/6 and P(C=1)>1/6. Thus, the simulation value P(C=1) = 1/3 may be reasonable, but that of P(C=0) may not. This discrepancy is of concern since **ITT** is the most sensitive to the distribution of true intake.

In study 6, where P(C) is heavily skewed towards high compliance (P(C=0)=.25, P(C=.5)=.25, and P(C=1)=.5), **ITT** has higher power than **ML**. Even in this instance, which likely illustrates a more favorable P(C) than is clinically relevant, **ML** yields the best estimates of P(Y|D). If anything, the values chosen for P(C) favor the performance of methods that believe compliance as measured.

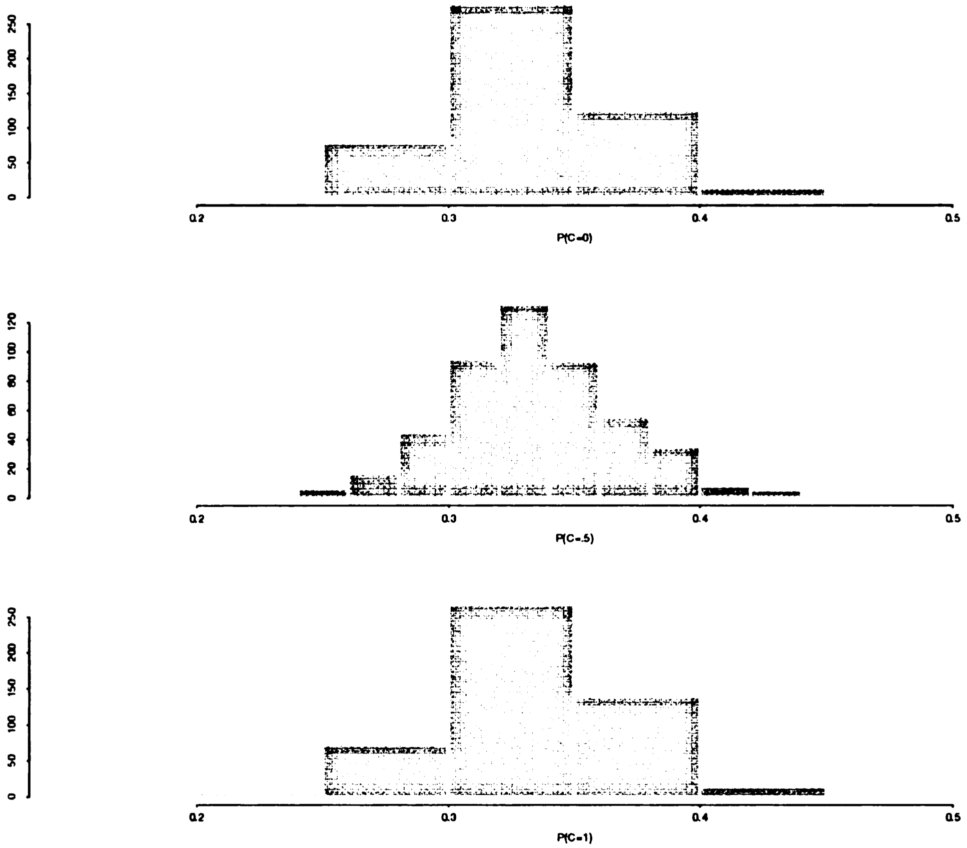


Figure 8.4. Histogram of P(C) in Study 1 Data. It is equally likely that a subject has C=0, C=.5, or C=1.

The prior distributions on P(Y|D) and P(C|C_Q) are used simply as an aid in comparing method performance. They allow all methods of analysis to return an estimate of exposure-response when data are sparse, and, consequently, simplify the task of evaluating relative method performance. After all, it is difficult to compare the performance of a method that only estimates exposure-response when it has rich data to a method that always returns an estimate of P(Y|D), regardless of data quality.

Here, the prior on P(Y|D) assumes no drug effect. The prior is always correct for P(Y|D=.5), but this is of little concern since this exposure-response estimate is not included in the evaluation of method performance. The prior P(C|C_Q) assumes no correlation between C_Q and C—an assumption that is incorrect for all but a few simulations in which cor(C_Q,C) is low. The **ML** method can be criticized for using this second source of prior information that no other method incorporates. However, the prior only aids analysis if it adds correct information. Otherwise, it does nothing or even hurts the analysis. Thus, **ML** is likely penalized by the incorrect prior on P(C|C_Q).

In practice, it is not necessary to use a prior with any method of analysis presented here. Of course, there may be instances in which one wants to use an informative prior. Sources of informative priors on P(C|C_Q) might be published data or information about a subject's compliance with another drug while sources of prior information on P(Y|D) include the drug effect measured in another clinical trial or the effect of a drug within the same therapeutic class.

8.2.2 Presentation of Results

One indirect consequence of the limited number of simulations performed is revealed in Figures 5.22, 5.23, and 5.24. The lines on the contour plots of (cor(C_Q,C), f_M, $\delta_{D=Dn}$) and (cor(C_Q,C), ρ, $\delta_{D=Dn}$) in study $1^{Dn=1}$ and study $5^{Dn=1}$ are wavy. This likely represents edge effects in smoothing—something influenced by a lack of data in extreme regions of parameter space.

Note that LHS only guarantees that points are spread out evenly in the parameter space it samples. The lack of points in certain regions of plotting space occurs because neither cor(C_Q,C) nor ρ are sampled by LHS. They are plotted because one has a more intuitive understanding of their values than of the LHS sampled parameter values which indirectly give rise to P(C_Q|C) and P(Y|D). Refer to the Appendix for an explanation of how P(C_Q|C) and P(Y|D) are simulated. The edge effects are expected to be less influential if more simulations are performed. The lack of data primarily makes it difficult to draw conclusions about parameter interactions. Because of this, few recommendations are made in this regard.

Estimation of P(Y|D) is performed on the logit scale because the logit of any probability has the desirable property of existing in unconstrained parameter space. All exposure-response probabilities, p, (on the 0-1 scale) are transformed to logits (ranging from -∞ to +∞) via equation 8.1.

$$x = \ln(p/(1-p)) \tag{8.1}$$

The logit is simply the log of the odds ratio.

Method performance is compared on the logit scale, as well. The Appendix explains how to compute various error metrics ($\delta_{D=0}$, $\delta_{D=Dn}$, Δ_R), but there is a need for a more intuitive understanding of their magnitude. Transforming the logit of a probability back to a probability aids in the understanding of this value. The antilogit is computed using

$$p = e^x/(1+e^x). \tag{8.2}$$

Given the optimal case in which there is zero logit error in the estimation of P(Y|D), the antilogit of that error is .5. That is, $p = e^0/(1+e^0) = ½$. Table 8.1 lists the mean absolute logit error, or the absolute error in the log odds ratio, in study 1. The corresponding antilogits for these values are listed in table 8.2.

Table 8.1. Mean Absolute Logit Error in Study 1.

	Mean Absolute Logit Error (Error in the log odds ratio)
ALL	.2654732
BA	.3778191
CD	.3604457
ITT	.6608587
BSR	.5188329
ML	.3190349

Table 8.2. Transformation of the Mean Absolute Logit Error in Study 1 to the Probability Scale.

	logit^{-1}(Mean Absolute Logit Error)
ALL	.5659812
BA	.593347
CD	.5891483
ITT	.6594533
BSR	.6268748
ML	.579089

Since the lowest value the antilogit of an absolute error can assume is .5, the results in Table 8.2 may be interpreted as showing how close each method would come to estimating a true value of p=.5. **ALL** estimates the probability best with p=.566, while **ITT** is the least accurate yielding p=.659. Because three probabilities are estimated (P(Y|D=0), P(Y|D=.5D_n), P(Y|D=D_n)), it seems confusing to represent the error as a value relative to p=.5. To provide a sense of scale, it makes more sense to compare each method's estimation precision to **ALL**'s estimation precision. Δ_R achieves this. Table 8.3 lists the mean Δ_R reported in study 1. They are computed by dividing the values in Table 8.1 by **ALL**'s value in Table 8.1. (See Figure 5.1 for the entire Δ_R distribution.) This transformation delivers the absolute error in the log odds ratio for each method relative to the absolute error in the log odds ratio for **ALL**.

Table 8.3. Mean Absolute Logit Error Relative to ALL in Study 1.

	Mean Absolute Logit Error Relative to ALL
ALL	1
BA	1.423191
CD	1.357748
ITT	2.489361
BSR	1.95437
ML	1.201759

For the sake of continuing the discussion of the intuitive interpretation of Δ_R, Table 8.4 lists the mean Δ_R reported in Table 8.3 in terms of its corresponding probability on the antilogit scale.

Table 8.4. Transformation of Mean Absolute Logit Error Relative to ALL in Study 1 to the Probability Scale.

	logit^{-1}(Mean Absolute Logit Error Relative to ALL)
ALL	.7310586
BA	.8058382
CD	.7953934
ITT	.9233926
BSR	.8759224
ML	.7688376

When the errors are computed relative to **ALL**, the value of the relative mean absolute logit error of comparison is 1, and the corresponding p is $p=\text{logit}^{-1}(1)$, or .731. Again, this probability does not have a literal interpretation. Comparing performance to **ALL** is a good idea, but one must be careful to interpret the plots of Δ_R as an illustration of relative method performance, not extract quantitative information on method performance from them. The results are presented in Chapters 5-7 in a manner that reflects this limitation.

The confidence regions around $\delta_{D=Dn}$ is determined by bootstrap. The procedure for computing it is explained in the Appendix. As with all bootstrapped confidence intervals, these "confidence regions" underestimate variance(Efron 1993; Mooney and Duval 1993). There is no reason to suspect that **ML** gains any advantage over other methods of analysis with respect to the bootstrap computation of confidence regions. However, one must be wary about interpreting the variance quantitatively.

Power is determined in a region of low power for all methods of analysis. This is done to reduce the number of computations required to compute power. Although there is no reason to believe that relative method performance will change with more simulations, these plots should only be interpreted qualitatively.

8.2.3 Are the Interesting Aspects of the Measurement Problem Preserved?

The critique of the parameter space explored and the representation of method performance suggests that the qualitative results presented in Chapters 5-7 are trustworthy. One can reasonably conclude that ML is the best method of analysis to use given data of the type analyzed in those studies. In this section, the entire simulation study design is scrutinized in order to determine the extent to which the results generalize to different types of data. The influence of model structure and assumptions on generality is discussed.

8.2.3.1 Model Structure

A simple clinical trial design is assumed in this investigation. Subjects are either assigned to drug or placebo. Drug exposure is a categorical variable that has a log linear causal relationship with one's success or failure. Compliance is measured, at most, once with each tool. Accurate compliance data are missing completely at random. The pattern of missingness is monotone (only C_M data are missing). True compliance, electronically monitored compliance, and self-reported compliance are unconfounded with response. There is no inter- or intra- individual variability in parameters.

Many of these simplifying design features are clinically relevant while others have little bearing on the problem at hand. It makes no difference, technically speaking, whether the one value of compliance used to determine exposure reflects a single observation or the mean of several compliance measurements. Likewise, the categorization of fractional compliance is general enough to extend to other metrics of intake. Although C is presented as the amount of drug taken, it could very well represent the timing of dose taking (subjects can either take doses perfectly on schedule, moderately on schedule, or at random time intervals), the frequency of drug holidays, or something else.

The assumptions that C_Q and C_M are unconfounded with response are justifiable. With the exception of drugs that impair an individual's ability to self-report compliance or use an electronic monitor, it is difficult to imagine a mechanism by which C_Q and C_M supply information about drug response independent of what they indicate about exposure. For example, a drug that affects a patient's memory may yield C_Q values that are causally related to Y. Likewise, an anti-arthritic drug which affects a patient's ability to use an electronically monitored medication bottle appropriately can give rise to C_M values that are causally related to Y.

One should be mindful of these exceptional cases when deciding whether the assumptions $P(Y|D_n,C,C_Q)=P(Y|D_n,C)$ and $P(Y|D_n,C,C_M)=P(Y|D_n,C)$ are reasonable. If such confounding exists, any method that uses compliance to determine exposure will yield biased estimates of the exposure-response relationship. If one tool is confounded with response, then the choice of analysis methods may be limited to those that do not use the confounded instrument. Note that in the measurement error framework, it is generally acceptable to assume nondifferential error when W (here: C_Q or C_M) is merely a mismeasured version of X (here: C)(Carroll 1995).

The model for drug exposure neglects to account for clearance and bioavailability. The clinical relevance of this simplification depends on the relative contribution of compliance and CL/F to exposure. The formulation is reasonable for drugs with little variability in CL and F. It may also be relevant for drugs with moderate variability in CL and F since compliance may range from 0-100%(Kass, Meltzer et al. 1986). In reality, one would include CL/F in a model for exposure (if the parameter was known) during the model selection process. If it has a greater impact on exposure than compliance, compliance adds little to an analysis, and the use of any compliance information, regardless of how it is measured, is inconsequential. Most importantly, however, since CL/F is just a scale term (in the sense that its impact on the estimation of exposure is not influenced by the method of determining compliance), this simplification has no bearing on the problem considered here. That is, variability in CL/F is not expected to affect the performance of one method of analysis more than another.

In the interest of minimizing computational time, this investigation focuses on categorical data. If one's performance metric is estimation time, then switching to continuous models may hurt ML. However, computation time is less of an issue when fitting a model to a single data set than in simulation studies where a large number of trials must be run. Note that one benefit of switching over to continuous models is that they may have fewer parameters than categorical models, and, consequently, handle sparse data better than the categorical formulation presented here. In that case, continuous models reduce the influence of prior information on estimation. Recall that the prior penalizes **ML** more than other methods of analysis, thus, the categorical formulation may present **ML** in its worst light.

Here, a monotone missing data structure is assumed. That is, C_Q and response data are complete and only C_M data are missing. In reality, more complicated missing data patterns arise. Both self-reported compliance and electronically monitored compliance data are missing in the data source for Chapter 7. However, the C_Q data are more abundant than C_M data—100 of Bangsberg and coworkers' 102 self-report assessments during the clinical trial are complete, while only 77 of 102 electronically monitored assessments are complete(Bangsberg, Hecht et al. 2000). The **ML** method as presented is unable to handle this pattern of missingness. To apply the method to such a data set, one may have to throw away records from subjects missing C_Q data. (This assumes one integrates over $P(C|C_Q)$. One might choose to parameterize the model such that $P(C)$ or $P(C_Q)$ is estimated. Doing so allows one to integrate over $P(C)$ or $P(C_Q)$ in subjects missing both C and C_Q.) In the case of the Bangsberg and coworkers' data set, it is likely that **ML** maintains its superior performance when the two records with incomplete C_Q data are dropped. Note that a lack of C_Q data is expected to negatively impact **BSR** and **BA**. Thus, the decision to analyze a monotone data set does not reflect a bias in favor of **ML**.

8.2.3.2 Assumptions

Although the majority of the design decisions are not expected to influence conclusions about method performance, two data analytic assumptions pose great challenges to the generality of ML.

The first troublesome assumption is that C_M is missing completely at random (MCAR). Under the design presented, the missingness of C_M is an investigator-determined parameter. In reality, C_M may be missing for other reasons. ML is expected to yield biased estimates of exposure-response if C_M's missingness depends on the value of C_M. However, even in this case, ML may yield less biased exposure-response estimates than other methods of analysis. One may predict whether ML yields reasonable estimates of exposure-response by determining the extent to which C_M is missing at random (MAR). Recall from the discussion in Chapter 2 that C_M data only have to be MAR, not MCAR, in order for ML to yield unbiased estimates of exposure-response.

One may determine whether the MAR assumption is reasonable by comparing the distribution parameters in subjects who have C_M data to those who do not (i.e. $P(C_Q,Y|M=0)$ vs. $P(C_Q,Y|M=1)$). If data are not MAR, one may bootstrap the data sets to determine if just a few subjects have data that is not MAR. Procedurally, one randomly samples the data with substitution to generate a new data set that is the same length as the original data set. Then one computes and compares some statistic on $P(C_Q,Y|M=0)$ and $P(C_Q,Y|M=1)$. This procedure is repeated many times. The overlap between the statistic for subjects with M=0 and M=1 should give some indication of whether data are MAR for many or just a few subjects.

The assumption that C is unconfounded with Y is of great concern, as well. It is not an assumption that can be supported using available data. There are instances in which the data suggest that confounding exists. For example, the clinical trial of cholestyramine revealed that good compliers to placebo have a better outcome than poor compliers with placebo(Efron and Feldman 1991). It is a drug taken chronically, thus, some have suggested that compliance is a surrogate for lifestyle factors that impact the subject's outcome. Regardless of whether one can imagine a mechanism by which compliance relates to response independent of exposure, critics of as-treated analyses can rightfully argue that confounding always may be present.

If true compliance is confounded with response, ML, as well as all other methods that use compliance information to determine exposure, yield biased estimates of exposure-response. However, the degree of bias may be less than that incurred by ignoring compliance information altogether and using ITT to estimate exposure-response. Since the method of determining exposure is not likely to have an interaction with confounding, there is no reason to suspect that confounding causes ML to lose its relative advantage over other methods that use compliance data as measured.

One weakness of this investigation is that it neglects to explore sensitivity to the assumptions of MCAR and nonconfounding of true compliance. Furthermore, the ML method may be helped by estimating an exact likelihood. The real world is more complicated than it has been presented here. Model selection should be less straightforward, particularly for this model-based method.

Clearly, this work leaves unanswered questions. However, it presents enough evidence to suggest that the ML method should be used to analyze calibration studies. The benefits outweigh potential risks.

8.2.4 Impact of this Work on Pharmaceutical Science

This is not the first proposal to use compliance data to determine exposure and estimate exposure-response from "natural dosing experiments"(Urquhart and Chevalley 1988). The idea that several tools can be used together to obtain a better estimate of compliance is not new, either(Jonsson, Wade et al. 1997; Liu, Golin et al. 2001). However, no one has approached compliance determination as a missing data problem or used a maximum likelihood model-based approach to determine intake. Others have investigated a technique for choosing the best measure of compliance among contenders(Jonsson, Wade et al. 1997) or evaluated single imputation approaches(Liu, Golin et al. 2001).

The model-based calibration of compliance information is the most general approach among those presented. It can easily be expanded to incorporate more complex clinical trial designs. And it can be extended to eliminate the need for assumptions such as C_M=C, ignorable missingness of C_M, and nonconfounding of compliance under certain assumptions.

This investigation is the most thorough comparison of various methods for using compliance data as measured. This task is helped by the method used to explore parameter space. Latin Hypercube Sampling is one that has not previously been used in PK/PD simulation studies. The results show that given self-reported compliance with some correlation between C_Q and C, it is better to use **BSR** than **ITT** to determine exposure-response. However, the benefit does not extend to power. If there are some C_M data available, and cor(C_Q,C) is greater than 0, it is better to use the best available measure of compliance to determine exposure than to throw away self-reported compliance. Most surprisingly, even when there is no correlation between C_Q and C_M, ML yields the best estimates of exposure-response.

In this study, self-reported compliance is labeled the more feasible and less accurate compliance monitoring tool. This nomenclature is based on information in the literature reporting on the challenges of measuring compliance using self-report tools. However, it does not escape our attention that self-reported intake can be more accurate than electronically monitored compliance in certain instances. Regardless, the analyses presented here are not contingent on self-reported compliance specifically being a poorer measure of compliance than electronic monitors. C_Q should be generically interpreted as a tool that measures compliance with more error than C_M. The important point is that the ML method as presented is recommended, provided that the data analyst correctly identifies which tool is superior.

8.3 References

Bangsberg, D. R., F. M. Hecht, E. D. Charlebois, A. R. Zolopa, M. Holodniy, L. B. Sheiner, J. D. Bamberger, M. A. Chesney and A. Moss (2000). “Adherence to Protease Inhibitors, HIV-1 Viral Load, and Development of Drug Resistance in an Indigent Population.” AIDS **14**(4): 357-366.

Burney, K. D., K. Krishnan, M. T. Ruffin, D. Zhang and D. E. Brenner (1996). “Adherence to Single Daily Dose of Aspirin in a Chemoprevention Trial. An Evaluation of Self-Report and Microelectronic Monitoring.” Archives of Family Medicine **5**(5): 297-300.

Carroll, R. J., Ruppert, D., and Stefanski, L.A. (1995). Measurement Error in Nonlinear Models. Great Britian, St. Edmundsbury Press.

Catania, J. A., D. Binson, J. Canchola and L. M. Pollack (1996). Effects of Interviewer Gender, Interviewer Choice, and Item Wording on Responses to Questions Concerning Sexual Behavior. Public Opinion Quarterly, University of Chicago Press. **60:** 345-375.

Efron, B. (1993). An Introduction to the Bootstrap. New York, Chapman & Hall.

Efron, B. and D. Feldman (1991). "Compliance as an Explanatory Variable in Clinical Trials." Journal of the American Statistical Association **86**: 9-22.

Ickovic, J. R. and A. W. Meisler (1997). "Adherence in AIDS Clinical Trials: A Framework for Clinical Research and Clinical Care." Journal of Clinical Epidemiology **50**(4): 385-391.

Jonsson, E. N., J. R. Wade, G. Almkvist and M. O. Karlson (1997). "Discrimination between Rival Dosing Histories." Pharmaceutical Research **14**: 984-991.

Kaplan, R. M. and H. J. Simon (1990). "Compliance in Medical Care: Reconsideration of Self-Predictions." Annals of Behavioral Medicine, Society of Behavioral Medicine **12**(2): 66-71.

Kass, M. A., D. Meltzer, M. Gordon, D. Cooper and J. Goldberg (1986). "Compliance with Topical Pilocarpine Treatment." American Journal of Ophthalmology **101**: 515-523.

Little, R. J. A. and D. B. Rubin (1987). Statistical Analysis with Missing Data. New York, John Wiley & Sons.

Liu, H., C. E. Golin, L. G. Miller, R. D. Hays, K. Beck, S. Sanandaji, J. Christian, T. Maldonado, D. Duran, A. H. Kaplan and N. S. Wenger (2001). "A Comparison Study of Multiple Measures of Adherence to HIV Protease Inhibitors." Annals of Internal Medicine **134**: 968-977.

McKay, M. D., R. J. Beckman and W. J. Conover (1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code." Technometrics **21**(2): 239-245.

Mooney, C. Z. and R. D. Duval (1993). Bootstrapping: A Nonparametric Approach to Statistical Inference. Newbury Park, Sage Publications.

Straka, R. J., J. T. Fish, S. R. Benson and J. T. Suh (1997). "Patient Self-Reporting of Compliance Does Not Correspond with Electronic Monitoring: An Evaluation Using Isosorbide Dinitrate as a Model Drug." Pharmacotherapy **17**(1): 126-132.

Turner, B. J. and F. M. Hecht (2001). "Improving on a Coin Toss to Predict Patient Adherence to Medications." Annals of Internal Medicine **134**(10): 1004-1006.

Urquhart, J. (1997). "The Electronic Medication Event Monitor. Lessons for Pharmacotherapy." Clinical Pharmacokinetics **32**(5): 345-356.

Urquhart, J. and C. Chevalley (1988). "Impact of Unrecognized Dosing Errors on the Cost and Effectiveness of Pharmaceuticals." Drug Information Journal **22**: 363-378.

Chapter 9: Future Directions

Abstract

Chapter 8 highlights the limitations of the methods and results presented in this report. Strategies for addressing these issues of generality and future elaborations of the methodology are presented in this section.

9.1 Introduction

Dosing guidelines are often developed via an iterative process that only begins during clinical trials. Of all drugs granted FDA approval between 1980 and 1999, 22% underwent significant post-marketing dose adjustment(s). Most often, the dose originally recommended in product labeling was too high(Cross, Lee et al. 2001).

Reducing dose over time can have dire consequences on a drug's pricing structure. More importantly, it jeopardizes patient safety. Incorrect dosing is estimated to be the 4th-6th leading cause of death in the United States. Each year, over 100,000 persons in the United States are killed by drugs taken as directed. This statistic excludes adverse drug reactions caused by errors in drug administration, noncompliance, overdose, drug abuse, and therapeutic failure(Lazarou, Pomeranz et al. 1998). While some consider this value an overestimate(Fremont-Smith, Kravitz et al. 1998), other research suggests that it may be an underestimate. Approximately 125,000 deaths per year have been attributed to noncompliance with cardiovascular drugs alone(Bond and Hussar 1991).

To some extent, learning to dose must be a trial-and-error process. Only a small sample of the population can, realistically, be studied during clinical trials. By chance, subjects with unusual kinetic parameters may not be administered the drug until it is marketed. The pharmaceutical industry, regulatory officials, and patient groups are interested in making drug development a more efficient process. Increasing the size and duration of clinical trials to get the dose right runs counter to this goal.

Yet, the contribution of two likely causes for misdosing can be reduced without increasing the number of subjects enrolled in a clinical trial. One method is to select the dose(s) admitted into confirmatory clinical trials using a scientific, rather than an empiric, approach by performing data analyses that maximize the information gained from early phase clinical trials. The second way to improve dosing guidelines is to analyze confirmatory clinical trial data via an as-treated analysis to determine the exposure-response relationship. Clinical trialists can identify an optimum dose, based on the experience of a large number of patients, not just recommend the dose that causes one to reject the null hypothesis of no drug effect.

Noncompliance during clinical trials may be used to the drug developer's advantage, then, if it is treated as something that gives rise to natural dosing experiments(Urquhart 1992; Urquhart 1993). The many challenges to this data analytic approach have been presented in this thesis. In the process of answering methodological questions regarding the use of compliance information in computing exposure, these simulation studies raise further questions. Approaches to answering them are presented in this chapter.

The length of this "to do" list illustrates the tremendous amount of work required before compliance data will be widely used in the development of dosing guidelines. But the need to develop safe and effective therapeutics more efficiently than via empirical approaches is too great to allow the amount of work to be an acceptable deterrent.

9.2 Further Simulation Study

9.2.1 Extensions of the Simulation Studies Presented

An alternative approach to carrying out simulations is to evaluate method performance using the closed form solution for each analysis method's estimation precision. Unfortunately, the prior distributions on P(Y|D) and P(C|C$_Q$) hinder one's ability to mathematically describe estimation precision in the studies presented. Analytic solutions are likely to be unattainable given more realistic models for compliance behavior, clinical trial design, clinical trial conduct, and patient pharmacokinetic parameters, as well.

A more thorough investigation of the influence of P(C$_M$|C), P(C), and P(Y|D) on method performance will increase confidence in the results presented. It may be useful to choose P(C$_M$|C) in a manner similar to that used to select P(C$_Q$|C). (Refer to the Appendix for a description of how P(C$_Q$|C) is selected at random). That is, $C_{M,M\leq C}$ can be averaged with $C_{M,M=C}$ and $C_{M,M\perp C}$ with randomly varying weights. Of course, for P(C$_M$|C), it makes sense to average in $C_{M,M\perp C}$ to a lesser extent than $C_{Q,Q\perp C}$ contributes to P(C$_Q$|C). Furthermore, since changes in technology may improve the accuracy of electronic monitoring tools, weighting of $C_{M,M=C}$, $C_{M,M\perp C}$, and $C_{M,M\leq C}$ should be such that the mean of the distribution of cor(C$_M$,C) is shifted upward compared to the distribution already investigated in sensitivity studies.

It was pointed out that the model for P(C) used may overestimate the fraction of subjects who are fully compliant and noncompliant but underestimate the probability of subjects who are moderately compliant. To address this issue, the clinically observed distribution of C$_M$ values, P(C$_M$=0)=1/6, P(C$_M$=.5)=2/3, and P(C$_M$=1)=1/6, should be used as P(C) in simulation studies. Additionally, one may take the distribution of compliance as determined via unannounced pill counts as a possible P(C).

The model for P(Y|D) should be altered such that it doesn't pivot around a fixed probability of success at D= .5D_n. A clinically relevant approach is to select several models for exposure-response from the literature. One may want to choose a drug causing a small difference in response, as is typical of long term outcomes such as the change in mortality conditional on exposure to cholesterol lowering drugs. This is of interest since compliance tends to decrease as a function of time and poor compliance is an issue with drugs taken for a long period of time(Cramer, Scheyer et al. 1990; Waeber, Leonetti et al 1999).

Among all of the criticisms presented in Chapter 8, the issues raised about power are the most computationally expensive to address. Estimation of power requires hundreds of simulations under each set of parameter values. It is more costly to make statements about power that are as general as the statements made about estimation precision. The discrepancy between power and estimation precision for **ITT** and **BA**, however, suggests that there is a need to explore power further. Power should be computed in many regions of parameter space, as when parameter values are sampled by LHS, rather than in a few places chosen to favor one method over another. The sensitivity of power to all assumptions of data simulation should be determined, as well.

To allow a fair comparison between all methods of analysis, the prior on P(C|C_Q) was uninformative. Although this conservative choice did not draw criticism, it would be interesting to repeat study 1, compute power, and perform sensitivity analyses using **ML** with a more informative prior on $P(C|C_Q)$. For example, θ_{2prior} can be based on data in the literature.

Further investigation into the influence of assumptions on method performance is needed. One may want to explore sensitivity to the assumption that C_M is missing completely at random. To determine the effect of nonignorable missingness, M must be simulated from a model in which assignment of the electronic monitor depends on a subject's C_M. For example, if subjects with low compliance are more likely to lose their electronic cap, $P(M=1|C_M)$ might be distributed such that $P(M=1|C_M=0)<$ $P(M=1|C_M=.5)< P(M=1|C_M=1)$. Because this parameter is one for which clinically relevant values can only be speculated, it should be varied widely using LHS.

Note that although **ML** is sensitive to the MCAR assumption, reasonable violations of the assumption are not expected to impact **ML**'s relative performance. **CD** is expected to suffer the most when $P(M|C_M) \neq P(M)$ since it already yields highly variable estimates of exposure-response due to having the fewest data. Since **BA**'s performance depends on C_M data, its $P(Y|D)$ estimates are expected to worsen, as well. Given that **ITT** and **BSR** yield extremely biased estimates of $P(Y|D)$, **ML** is likely to maintain its competitive advantage over these methods.

To explore the consequence of having an electronically monitored value of compliance depend on the subject's response (C_M is MAR), M must be simulated conditional on Y. For example, if subjects who respond poorly to anti-arthritic drugs are less likely to use the medication cap appropriately, the availability of C_M data might be distributed such that $P(M=1|Y=0) < P(M=1|Y=1)$. This parameter is one for which there is no information in the literature, so it should also be varied widely using LHS. To explore sensitivity to the assumption that the missingness of C_M data depends on both Y and C_M, M should be simulated from the model $P(M|C_M,Y)$. Given the number of possible models for $P(M|C_M,Y)$, one may just want to focus on a few extreme cases. When developing simulation models to investigate sensitivity to the assumption that C_M is MCAR, one should be sure to maintain a random component in $P(M|C_M)$, $P(M|Y)$, and $P(M|C_M,Y)$ if the underlying study design is a double sampling scheme in which some patients neglect to return the electronically monitored cap conditional on their C_M and/or Y.

By simulating C_Q and C_M conditional only on C, one implicitly assumes that error in C_Q is independent of error in C_M and vice versa. Depending on which compliance measuring tools are represented by C_Q and C_M, it is more or less difficult to imagine a mechanism by which this assumption is true. One can easily envision how error in pill counted compliance and error in self-reported compliance may be correlated. A subject who dumps pills may be more likely to lie about his intake. It is more difficult to imagine a relationship between the error in electronically monitored compliance and self-reported compliance. There seems to be no reason why a subject who removes all of his doses for the day at one time (i.e. has compliance underestimated by C_M) is more likely to overestimate his compliance when asked about drug intake. Depending on the application of the methodology, one may want to investigate method sensitivity to this assumption.

The exposure model presented does not consider the influence of clearance and bioavailability in the calculation of D. Although this is a gross simplification, it may be a necessary simplification in data analysis as CL/F may be unknown in subjects enrolled in a confirmatory clinical trial. To explore the relative impact of variability in compliance and variability in pharmacokinetic parameters on exposure, one may simulate exposure from a model including the effect of both factors, but compute exposure as in the studies presented. Several clinically relevant values of CL/F should be selected. Variability in CL/F is not expected to hurt relative method performance, however, as all methods that compute exposure will be affected by CL/F equivalently, and **ITT** is already very biased.

Given the level of concern in the statistical literature over the potential confounding of compliance with response, sensitivity to the assumption of nonconfounding may be addressed in concert with the measurement issue. To do this, simulation of Y should be carried out via a model in which C influences Y in two ways. For example, P(Y|D,C) may be formulated using a combination of an E_{max} model and a linear function of C, as in equation 9.1.

$$Y = \{[E_{max}*D(C,D_n)]/[D_{50} + D(C,D_n)]\} + \alpha C \tag{9.1}$$

This model requires specifying a value for α. Since it is unknown clinically, α should be varied widely. One may speculate on the value of α by reviewing the results of clinical trials of drugs treating long-term outcomes, which may be influenced by lifestyle factors. For example, one may want to take the average difference in response for placebo compliers and placebo noncompliers in the clinical trial of cholestyramine(Program 1984) as an estimate of α.

Some consider the potential for confounding to be an unequivocal argument against using an as-treated analysis(Lee, Ellenberg et al. 1991). They argue that **ITT**'s bias is acceptable because it is conservative. This may be true in hypothesis testing mode, however, it is likely anticonservative if it is used to develop dosing guidelines. A compliance average response may cause those who take the full amount of drug and develop toxicity to appear to be outliers, rather than the norm at the given dose. Consequently, doses recommended on the basis of an **ITT** analysis may be higher than necessary.

Before rejecting an as-treated analysis as a method for estimating the exposure-response relationship in the face of possible confounding, one should evaluate the possible resulting magnitude of bias in exposure-response estimates. Under the conditions explored, it is likely that **ML** maintains its superior performance even if confounding is present since all as-treated analysis methods are affected by confounding and **ITT** is very biased.

Note also that if only one dose is tested in a clinical trial, the **ITT** analysis is unable to deliver an estimate of exposure-response.

9.2.2 Explore More Relevant Models

To promote an intuitive understanding of the analysis methods, some clinical realism was forfeited in this investigation. With this foundation in place, the next step is to consider complexities beyond saturated models for categorical variables. Here, parametric models of continuous variables are explored. Since continuous models involve integration, one should expect an increase in computation time. However, the analysis of continuous data may allow for one simplification of the methods used in analyzing categorical data. The prior distributions on $P(Y|D)$ and $P(C|C_Q)$ may no longer be necessary if the model for continuous data is identifiable with fewer data.

The first level of complexity one should add is a more clinically relevant model for exposure-response. The E_{max} model can easily replace the logistic model explored. The model for drug exposure can be made more clinically relevant by using both compliance and pharmacokinetic data.

One may adopt a population pharmacokinetic approach to estimating the exposure-response model by treating compliance as something measured via a sparse sampling design. As with the simple scheme presented in this report, neglecting to consider the contribution of patient compliance to variability in exposure is a source of error in estimating the exposure-response model. The question posed is whether the error in compliance measurement causes more bias in exposure-response estimates than ignoring it altogether. If it helps, which particular method of analysis performs best? How does performance relate to CL/F and study design?

Of course, dose, compliance, and response data from just about any trial design can be analyzed in the scheme presented by transforming the data into discrete values. But categorizing data from a longitudinal study may waste information (signal). To preserve the richness of time varying responses, one may choose to model compliance and response dynamically. If one is willing to investigate the compliance measurement problem using dynamic models, a number of interesting questions can be explored.

A Markov Chain (MC) model has been used to parameterize the entire time series of compliance data as measured via an electronic monitor(Girard, Blaschke et al. 1998). The model has not been fit to the time series of self-reported compliance data in patient diaries. It would be interesting to compare the parameter estimates computed on both self-reported and electronically monitored compliance data to calibrate C_Q to C_M. In contrast to the model for compliance used in this report, the MC model offers a multivariate description of intake behavior. Despite the increase in the number of model parameters, the MC model may reveal similarities between certain characteristics of C_Q and C_M data that makes it easier to calibrate than via the nonparametric approach presented. Girard and coworkers demonstrate that truncating electronic records of dosing to several previous half lives can reduce the amount of computational time required for the MC model(Girard, Sheiner et al. 1996). Of course, this simplification depends upon whether the system is linear.

One may focus on capturing one particular aspect of the time varying nature of compliance—the observation that compliance decreases with time. Since compliance tends to plateau, an inhibitory Emax type model may describe its value. This is important to consider in the calibration framework since this formulation may impact the ability to calibrate one tool to another. Different tools may vary in sensitivity to detect the change in compliance over time.

Dynamic models allow one to investigate the impact of nonlinearity in pharmacokinetics on the determination of exposure. The relative importance of clearance and bioavailability versus compliance in determining exposure may vary with time if a drug exhibits nonlinear kinetics. This is important to address in the compliance measurement error framework since inaccurate compliance measurement may mask the contribution of compliance to exposure. Given that few drugs are admitted into clinical trials with nonlinear kinetics, however, this may not the most important issue to explore.

In contrast, it is clinically relevant to explore dynamic models for response since they allow one to describe rebound and withdrawal. Poor compliance has been implicated in rebound hypertension due to a drop in β-blocker concentration(Urquhart 1997). Noncompliance with Paxil® is thought to elicit withdrawal effects(Kehoe 2001). The measurement of compliance is important to consider with respect to these problems as inaccuracy in compliance may make it more difficult to discover rebound and withdrawal effects.

Dynamic models become prohibitively challenging to estimate and lose their intuitive quality as more parameter interactions are incorporated. When exploring dynamic models, one should not lose sight of the measurement question. Additionally, one should focus on the most clinically relevant relationships. Refer to Figure 9.1 for an aid in deciding which interactions may be important to investigate. It illustrates various mechanisms by which compliance, CL/F, exposure, and a subject's outcome may cause or be subject to feedback.

A hypothetical mechanism for the process illustrated by Arrow #1—showing that Y influences CL/F—is a case in which a lack of immunosuppressive efficacy causes rejection of the organ for eliminating the immunosuppressive. One can imagine several examples of the process illustrated by Arrow #2—exposure changes CL/F. Saturation of gut or liver metabolic enzymes can cause changes in a drug's bioavailability and/or clearance. A drug, which increases urination, will have increased clearance as a function of drug exposure if it is primarily renally eliminated (assuming significant tubular reabsorption does not occur). Arrow #3 signifies instances in which a subject's outcome influences compliance. Response to an anti-arthritic drug may affect how well a patient can administer doses via an electronically monitored pill dispenser. Additionally, compliance with drugs treating psychiatric disorders may, hypothetically speaking, be

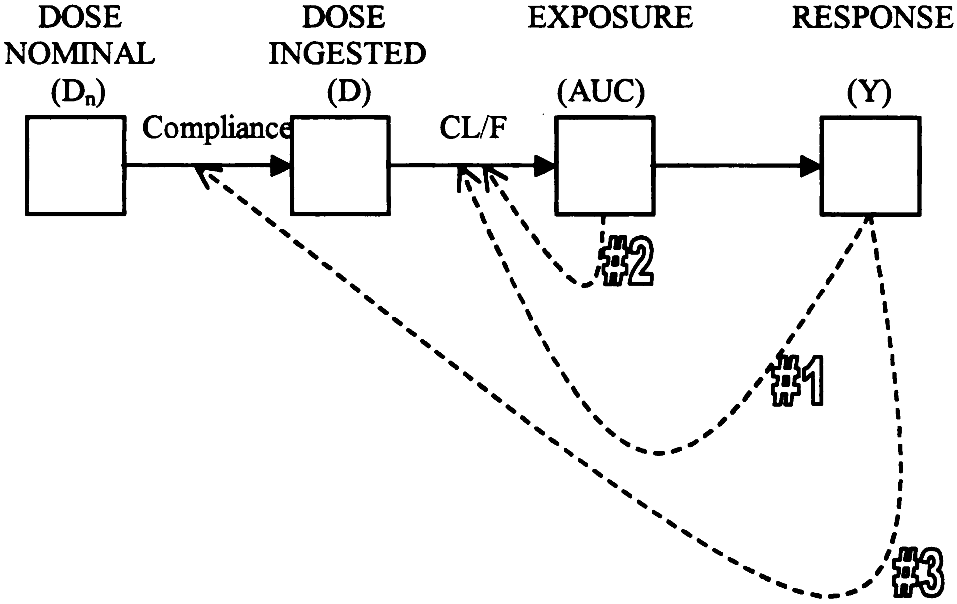


Figure 9.1. Schematic Representation of the Feedback of Response on CL/F, Exposure on CL/F, and Response on Compliance.

influenced by a patient's mental outcome. Drugs treating Alzheimer's disease can affect how well one remembers to take one's medication.

Note that only the outcome of interest is considered in this discussion of Figure 9.1. The impact of other responses on compliance, such as toxicity, are not mentioned here as this would require a model for multiple outcomes. Consider the example of how a favorable response to antibiotics causes reduced compliance in some patients. One may model this interaction as having the pharmacodynamic response influence some psychological factor which influences compliance.

In summary, Figure 9.1 illustrates that it is surprisingly difficult to imagine clinically relevant scenarios by which interactions between AUC, CL, and Y occur. Thus, it was reasonable to exclude consideration of them in the simple formulation presented in this report. Given the level of complexity possible in investigating compliance measurement, it is important to start with a conceptual understanding, as in the simple model presented, of how the methods of analysis perform. Of course, the simplification is only useful if it generalizes, qualitatively, to the complex case.

9.2.3 Explore Other Data Analysis Methods

The goal of this project is to compare a new data analytic approach with those used in practice. Therefore, only one model-based method of analysis (ML) is compared to various methods for using the data as measured. In addition to examining the performance of ML and other existing approaches in more complicated situations, if such investigation reveals additional problems, the next step would be to improve the methods so they can handle more complicated situations. Additionally, one may evaluate other model-based approaches.

In Chapter 6, it was shown that **ML** is sensitive to the assumption that C_M=C (***A3***). It is sensitive to ***A3*** because the likelihood model assumes C_M is a perfect measure of C. **ML** and all methods that use C_M data could be changed to allow the estimation of the error term in C_M. Likewise, error in C_Q can be modeled in the methods that use self-reported compliance data. Of course, this requires more than one measure of C_M and C_Q to be taken or for information on error in compliance measurement to be available from another source of data.

One may, theoretically, change **ML** by proposing a missing data mechanism, $P(M|Y,C_M)$, in order to allow for unbiased estimation of exposure-response when C_M is not MCAR. Of course, this approach requires further assumptions that are untestable on the data set at hand. The problem of informative missingness is difficult to get around!

Since **ML** is sensitive to the assumption that $P(C_Q|C,M)=P(C_Q|C)$, one may change **ML** to estimate a different $P(C|C_Q)$ in subjects that are electronically monitored and subjects that are not. To be identifiable, one would have to state a model for the relationship between $P(C|C_Q,M=0)$ and $P(C|C_Q,M=1)$. This, too, requires more untestable assumptions.

If dynamic models are used, one may incorporate terms in **ML** allowing for estimation of the feedback of response on compliance and rebound effects due to a lapse in drug intake.

In addition, to investigate the challenges of compliance measurement using more clinically relevant models for patient variables, more realistic models for trial conduct can be entertained. Monotonic patterns of missingness are more the exception than the rule in conducting clinical trials. ML may be changed to handle nonmonotonicity in compliance measured using two or more tools by modeling the relationship between several tools. Once three tools are used, one no longer must assume that one method measures compliance accurately for the problem to be identifiable. Under the "latent variables analysis" framework, identifiability is still possible when two of three methods are biased and the other is an unbiased but noisy measure of the predictor variable(Dearcangelis 1993; Kaaks 1994; Heckman and Vytlacil 1999). ML becomes more computationally difficult as the number of missing variables integrated over increases. However, the methods that use compliance as measured will also be "penalized" by this design since they now do not have the advantage of having a complete data set. Of course, losing data is likely a more significant concern than computational difficulty.

Horton introduces other forms of the maximum likelihood method that can be explored(Horton and Laird 2001). Additionally, a Bayesian approach, where all parameters have prior distributions on their values and the entire posterior density is estimated, may be investigated. An estimate of the posterior density may be necessary if loss functions are used in the computation of performance. The instrumental variables (IV) approach is an interesting, yet simple, approach for simultaneously addressing confounding and noncompliance.

The idea of IV is that one can tease out the causal relationship between an unknown covariate, X, and the response, Y, by quantifying the relationship between a parameter (an instrument, T) that affects X, but not Y. In the instrumental variables framework, we assume X is the unknown exposure (AUC), W is the faulty measure of exposure (compliance), and T is a patient's pharmacokinetic parameters (CL/F). Conceptually, CL/F is something that can cause patients with different levels of compliance to have the same exposure to drug irrespective of response. To be considered an instrument, T must satisfy three conditions; it must be (1) correlated with X, (2) independent of W given X, and (3) independent of Y given X. In terms of the notation used here, CL/F must be (1) correlated with AUC, (2) independent of C given AUC, and (3) independent of Y given AUC.

Knowledge of pharmacokinetics tells us that conditions (1) and (2) are true. For condition (3) to be true, CL/F must only drive response via exposure. One can imagine mechanisms by which (3) is violated. For instance, if low clearance is indicative of organ failure, then CL/F may indicate something about response in addition to what it tells about drug exposure. One should keep this limitation in mind when applying the approach to a particular problem.

When choosing other analysis methods to investigate, the multiple imputation approach should be given serious consideration. This method of analysis is not investigated in this report because multiple imputation parameter estimates approach ML. However, as the models for P(Y|D), clinical trial conduct, and compliance increase in complexity, ML will become more computationally time consuming and multiple imputation may become a more appealing option. Regression calibration, however, is not used in this report and is not advocated for use in estimating probabilities because it can yield impossible values—probabilities greater than 1(Selen 1986).

9.2.4 Compare Performance With Respect to Other Metrics

In this thesis, method performance is quantified in terms of error in exposure-response estimates and power. One may want to define efficiency in terms of more intuitive and/or practical metrics. To aid in the design of clinical trials, one may want to determine whether it is more cost effective to add a patient with self-reported compliance data only or to add someone with C_Q and C_M. As model-based methods become more complex, it will be increasingly important to compute performance with respect to computational time.

The methods could have different relative performance with respect to these performance metrics. Although it is not presented as a contending method of analysis, **ALL**'s performance is investigated in this report. In all but the investigation of sensitivity to the assumption that C_M=C, **ALL** is the best method of analysis. It is not presented as a contender because it is assumed that **ALL** is an impossible design. Realistically, the decision about whether **ALL** is doable depends on how much it "costs" to collect C_M data, how good C_Q data are, the size of the study, etc.

9.3 Methodologic Development Helped by Learning More About Compliance

Model building is an iterative process. Models improve as more experimental data become available. Models, in turn, can be used to inform the experimentalist about what data should be collected. Here, recommendations are offered regarding what data can improve the data analytic methodology.

One should determine experimentally, to the extent possible, if the assumptions of data analysis are reasonable. For example, to investigate the assumption that the accuracy of self-reported compliance is not influenced by the presence of an electronic monitor, one might compare $P(C_Q|C_M)$ for subjects that are knowingly monitored to those who are not informed that an electronic chip records when the pill bottle is opened. It would be interesting to determine if compliance is a baseline covariate as is often assumed; see (Efron and Feldman 1991; Sheiner and Rubin 1995)or an outcome. Via the analysis of a crossover study in which compliance is measured, one can determine whether individuals have the same compliance behavior regardless of the treatment assigned.

The challenge of confounding can be reduced by knowing a patient's compliance at the outset of a clinical trial (assuming, of course, that it is a baseline covariate). Ideally, the search for predictors of compliance will yield measurable covariates enabling investigators to stratify on (baseline) compliance before running a trial. Perhaps as more sophisticated models for drug intake patterns are developed—such as the Markov model which is parameterized in terms of the probability of taking a particular number of doses during the current dosing event given the number of doses taken during the previous dosing event—a metric of compliance having individual predictors will emerge.

To evaluate the magnitude of confounding effects, one can validate exposure-response estimates computed using an as-treated analysis by running studies in which "noncompliance" is randomly assigned(Chowdhury, Joshi et al. 1980; Wang, Shi et al. 1982; Landgren and Diczfalusy 1984; Landgren and Csemiczky 1991; Vaur, Dutrey-Dupagne et al. 1995).

9.4 Closing remarks

Patient compliance has been referred to as the ultimate barrier to drug delivery(Urquhart 1989) because it sets the upper limit on drug exposure. Wide inter- and intra- individual variability in compliance indicates that it is a significant barrier to drug delivery, as well(Cramer, Mattson et al. 1989; Cramer, Scheyer et al. 1990). Despite its important role in therapeutics, the U.S. Food and Drug Administration does not require that the relationship between compliance and response be provided in all drug labeling(Peck 1999).

The intention-to-treat causal estimator is necessary for regulatory purposes. Neglecting compliance information in the analysis of clinical trials is viewed by many as a necessary, albeit, conservative approach since compliance is, at least at present, impossible to predict and difficult to measure. One must be mindful, however, that the fact of efficacy, as determined via an intention-to-treat analysis may be causal, but the degree of efficacy, may not be generalizable. By ignoring compliance information, the **ITT** estimator yields an estimate of the treatment effect for some average intake in the group studied in the clinical trial.

If the average intake in the study sample is representative of compliance in the population, the **ITT** estimate is a valid estimator of treatment effect for that population. But the nature of compliance poses a serious threat to the use of the **ITT** estimate for dosing purposes. Compliance may be adjusted at any time. Changing a formulation to an implant or dosing the drug in a hospital setting, for example, ensures that all subjects get full exposure to drug. Compliance may change with the prescribed regimen, as well. As most drug developers aim to make drug products employing once-a-day regimens, compliance for an initially multiple times per day drug is likely to improve over the drug's lifetime. Furthermore, a given individual's compliance may decrease with time on treatment, so a change in prescription duration may affect patient compliance. An ITT and as-treated analysis should be performed(Feinstein 1991).

Compliance should be thought of as providing a necessary estimate of precision in dosing guidelines. Rather than using the results of an intention-to-treat analysis as the ultimate guide to dosing, ITT's estimates of exposure-response should be interpreted as equivalent to or less than method effectiveness. Likewise, an as-treated analysis can be considered as providing an estimate of exposure-response that is equivalent to or greater than method effectiveness. Thus, dosing guidelines based on ITT are likely an upper limit and dosing guidelines based on an as-treated analysis represent a lower limit.

The results presented do not render models that use compliance as measured obsolete. While ML is more useful for determining exposure-response from clinical trials, it fails in one clinical application. If a clinician wants to compute a patient's expected response given compliance of any value, a plug in type model needs to be provided in the drug's package insert. The investigation presented in this report helps in deciding which model or models to report in labeling. Given the practicalities of compliance measurement in medical practice, models for all possible types of compliance measures that may be available clinically should be supplied. Depending on the availability of compliance information, the clinician will use the one that is most pertinent. The investigation presented here allows one to recommend one model over another in the instance that multiple measures are clinically available.

In theory, data analyses using compliance information can impact how efficiently one learns to dose drugs. Rather than running clinical trials where poor compliance muddles efficacy, one could diagnose dosing and compliance issues before giving up on a drug(Kastrissios and Blaschke 1997). However, it is difficult to find an example of how noncompliance with investigational drug regimens cause an effective drug to fail to demonstrate efficacy. Kastrissios and Blaschke (1997) argue that a survey of nine New Drug Applications (NDA) approved by the Food and Drug administration between 1994 and 1995 suggests the impact of noncompliance on the results of clinical trials. For each NDA, 1486-13026 subjects participated in 16-49 studies. An average of 25% of the studies failed to demonstrate efficacy (range: 9%-65%). FDA approval was based on 3-14 studies.

Thinking of compliance measurement as existing between the missing data and measurement error frameworks raises an important general issue about confirmatory clinical trial data analysis. Missing data is often viewed as a more difficult problem to solve than measurement error. Investigators throw away records from subjects with missing responses. Yet one rarely discards all of a patient's record because some variable cannot be measured precisely.

Fortunately, modeling missing data is not unprecedented in the analysis of clinical trials. Modeling of responses has been used to handle drop out during data analysis(Sambol and Sheiner 1991; Sheiner, Beal et al. 1997). Additionally, simple changes to study design can be performed to circumvent data analytic challenges caused by drop out(Sheiner and Rubin 1995). Investigation of methods for handling responses that fall below the quantification limit is an area of active research(Beal 2001).

Consideration of these analyses should be the rule, rather than the exception. After all, it is the ethical responsibility of the clinical trialist to extract as much information as possible about drug effect from available data.

Although imperfect compliance may allow one to tease out the exposure-response relationship from confirmatory clinical trials, this thesis should not be interpreted as advocating for poor compliance when testing drugs. This study design is dangerous for trials comparing a treatment to an active control since drugs with equivalent efficacy could be made to perform differently through differences in the distribution of compliance(Urquhart 2001). Given a choice, one should aim to have perfect compliance in clinical trials. Clinically, one would never raise dose to compensate for poor compliance, one would raise compliance.

Patient compliance is often blamed for the lack of drug efficacy. However, compliance is not likely to be taken seriously by patients until it receives the attention it deserves in package inserts. The recent Paxil lawsuit and a letter to the editor illustrates that patients want to be given specific warnings about side effects if they don't take drug(Kehoe 2001). Patients need to be encouraged to take the drug on schedule to get the desired response.

Patient compliance is an unusual covariate. It is not correlated with gender, age, socioeconomic status, race, education, or any other baseline covariate(Lerner, Gulick et al. 1998; Chesney 2000; Wright 2000). Yet it varies widely between and within patients. Therefore, it would benefit a wide range of people if it were investigated more seriously during drug development.

9.5 References

(1998). “About Once a Week, I Forget to Take My Blood Pressure Pill. What Should I Do When That Happens? Take an Extra One the Next Day?” Harvard Heart Letter **9**(3): 8.

Beal, S. L. (2001). “Ways to Fit a PK Model with Some Data Below the Quantification Limit.” Journal of Pharmacokinetics and Pharmacodynamics **28**(5): 481-504.

Bond, W. S. and D. A. Hussar (1991). “Detection Methods and Strategies for Improving Medication Compliance.” American Journal of Hospital Pharmacy **48**(9): 1978-1988.

Chesney, M. A. (2000). “Factors Affecting Adherence to Antiretroviral Therapy.” Clinical Infectious Diseases. Supplement. **30**: S171-S176.

Chowdhury, V., U. M. Joshi, K. Gopalkrishna, S. Betrabet, S. Mehta and B. N. Saxena (1980). “'Escape' Ovulation in Women Due to the Missing of Low Dose Combination Oral Contraceptive Pills.” Contraception **22**(3): 241-247.

Cramer, J. A., R. H. Mattson, M. L. Prevey, R. D. Scheyer and V. L. Ouellette (1989). “How Often Is Medication Taken as Prescribed? A Novel Assessment Technique.” Journal of the American Medical Association **261**(22): 3273-3277.

Cramer, J. A., R. D. Scheyer and R. H. Mattson (1990). “Compliance Declines between Clinic Visits.” Archives of Internal Medicine **150**(7): 1509-1510.

Cross, J. T., H. K. Lee, J. S. Nelson, C. V. Grudzinskas and C. C. Peck (2001). One in Five Marketed Drugs Undergoes a Dosage Change: 1980-1999. Abstract at the

Annual Meeting of the American Society for Clinical Pharmacology & Therapeutics, Orlando, Clinical Pharmacology and Therapeutics, P63.

Dearcangelis, G. (1993). “Structural Equations with Latent Variables.” Journal of Applied Econometrics **8**(1): 111-113.

Efron, B. and D. Feldman (1991). “Compliance as an Explanatory Variable in Clinical Trials.” Journal of the American Statistical Association **86**: 9-22.

Feinstein, A. R. (1991). Intention-to-Treat Policy for Analyzing Randomized Trials: Statistical Distoritions and Neglected Clinical Challenges. Patient Compliance in Medical Practice and Clinical Trials. J. A. Cramer and B. Spilker, Eds. New York, Raven Press: 359-370.

Fremont-Smith, K., G. R. Kravitz, T. Bush, R. Hanzlick, P. B. Baker, G. M. Hutchins, K. Hui, J. Lazarou, B. H. Pomeranz, P. N. Corey and D. W. Bates (1998). “Adverse Drug Reactions in Hospitalized Patients.” Journal of the American Medical Association **280**: 1741.

Girard, P., T. F. Blaschke, H. Kastrissios and L. B. Sheiner (1998). “A Markov Mixed Effect Regression Model for Drug Compliance.” Statistics in Medicine **17**(20): 2313-2333.

Girard, P., L. B. Sheiner, H. Kastrissios and T. F. Blaschke (1996). “Do We Need Full Compliance Data for Population Pharmacokinetic Analysis?” Journal of Pharmacokinetics and Biopharmaceutics **24**(3): 265-282.

Heckman, J. J. and E. J. Vytlacil (1999). “Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects.” Proceedings of the National Academy of Sciences of the United States of America **96**(8): 4730-4734.

Horton, N. J. and N. M. Laird (2001). "Maximum Likelihood Analysis of Logistic Regression Models with Incomplete Covariate Data and Auxiliary Information." Biometrics **57**(1): 34-42.

Kaaks, R., Riboli, E., Esteve, J., Van Kappel, A.L., Van Staveren, W.A. (1994). "Estimating the Accuracy of Dietary Questionnaire Assessments: Validation in Terms of Structural Equation Models." Statistics in Medicine **13**: 127-142.

Kastrissios, H. and T. F. Blaschke (1997). "Medication Compliance as a Feature in Drug Development." Annual Review of Pharmacology and Toxicology **37**: 451-475.

Kehoe, W. A. (2001). "SSRI Discontinuation." The Prescriber's Letter.

Landgren, B. M. and G. Csemiczky (1991). "The Effect of Follicular Growth and Luteal Function of "Missing the Pill". A Comparison between a Monophasic and a Triphasic Combined Oral Contraceptive." Contraception **43**(2): 149-159.

Landgren, B. M. and E. Diczfalusy (1984). "Hormonal Consequences of Missing the Pill During the First Two Days of Three Consecutive Artificial Cycles." Contraception **29**(5): 437-446.

Lazarou, J., B. H. Pomeranz and P. N. Corey (1998). "Incidence of Adverse Drug Reactions in Hospitalized Patients: A Meta-Analysis of Prospective Studies." Journal of the American Medical Association **279**(15): 1200-1205.

Lee, Y. J., J. H. Ellenberg, D. G. Hirtz and K. B. Nelson (1991). "Analysis of Clinical Trials by Treatment Actually Received: Is It Really an Option?" Statistics in Medicine **10**: 1595-1605.

Lerner, B. H., R. M. Gulick and N. N. Dubler (1998). "Rethinking Nonadherence: Historical Perspectives on Triple-Drug Therapy for HIV Disease." Annals of Internal Medicine **129**(7): 573-578.

Peck, C. C. (1999). Non-Compliance and Clinical Trials: Regulatory Perspectives. Drug Regimen Compliance: Issues in Clinical Trials and Patient Management. J. M. Metry and U. A. Meyer, Eds. Chichester, John Wiley & Sons: 97-102.

Program, L. R. C. (1984). "The Lipid Research Clinics Coronary Primary Prevention Trial Results, Parts I and Ii." Journal of the American Medical Association **251**: 351-374.

Sambol, N. C. and L. B. Sheiner (1991). "Population Dose Versus Response of Betaxolol and Atenolol: A Comparison of Potency and Variability." Clinical Pharmacology and Therapeutics **49**(1): 24-31.

Selen, J. (1986). "Adjusting for Errors in Classification and Measurement in the Analysis of Partly and Purely Categorical Data." Journal of the American Statistical Association **81**: 75-81.

Sheiner, L. B., S. L. Beal and A. Dunne (1997). "Analysis of Nonrandomly Censored Ordered Categorical Longitudinal Data from Analgesic Trials." Journal of the American Statistical Association **92**(440): 1235-1244.

Sheiner, L. B. and D. B. Rubin (1995). "Intention-to-Treat Analysis and the Goals of Clinical Trials." Clinical Pharmacology and Therapeutics **57**(1): 6-15.

Urquhart, J. (1989). Non-Compliance: The Ultimate Absorption Barrier. Novel Drug Delivery and Its Therapeutic Applications. L. F. Prescott and W. S. Nimmo, Eds. Chichester, John Wiley & Sons: 127-137.

Urquhart, J. (1992). “Ascertaining How Much Compliance Is Enough with Outpatient Antibiotic Regimens.” Postgraduate Medical Journal. Supplement. **68**: S49-S58; discussion S59.

Urquhart, J. (1993). “Variable Patient Compliance in Ambulatory Trials--Nuisance, Threat, Opportunity.” The Journal of Antimicrobial Chemotherapy **32**(4): 643-649.

Urquhart, J. (1997). “The Electronic Medication Event Monitor. Lessons for Pharmacotherapy.” Clinical Pharmacokinetics **32**(5): 345-356.

Urquhart, J. (2001). “Demonstrating Effectiveness in a Post-Placebo Era.” Clinical Pharmacology and Therapeutics **70**(2): 115-120.

Vaur, L., C. Dutrey-Dupagne and J. Boussac (1995). “Differential Effects of a Missed Dose of Trandolapril and Enalapril on Blood Pressure Control in Hypertensive Patients.” Journal of Cardiovascular Pharmacology **26**: 127-131.

Waeber, B., G. Leonetti, R. Kolloch and G. T. McInnes (1999). “Compliance with Aspirin or Placebo in the Hypertension Optimal Treatment (HOT) Study.” Journal of Hypertension **17**(7): 1041-1045.

Wang, E., S. Shi, S. Z. Cekan, B. M. Landgren and E. Diczfalusy (1982). “Hormonal Consequences of "Missing the Pill".” Contraception **26**(6): 545-566.

Wright, M. T. (2000). “The Old Problem of Adherence: Research on Treatment Adherence and Its Relevance for HIV/AIDS.” AIDS Care **12**(6): 703-710.

APPENDIX

A.1 Simulation Design

The contending analysis methods are evaluated through simulation. Data are generated from known models of dose assignment, compliance, and exposure-response and then the simulated data are analyzed by all methods under consideration. With simulated data, the investigator knows the true drug effect causing patient response—enabling computation of estimation error. To focus on the compliance measurement issue, compliance data are treated as if they are unconfounded with response.

A.1.1 Simulation Design For Each Study

For generality, performance of the data analysis methods is determined over a wide range of investigator determined design parameters (N, f_M) and uncontrollable ($p(C_Q|C)$, $P(Y|D)$) parameter values. To do this efficiently, a multidimensional stratified sampling technique known as Latin Hypercube Sampling (LHS) is used(McKay, Beckman et al. 1979; Iman and Helton 1988; McKay, Beckman et al. 2000). In LHS sampling, each range of the X_k parameters in sample space, S, is divided into N_{lhs} strata of equal marginal probability $1/N_{lhs}$. For each simulation of N_{lhs} total simulations, parameter values are drawn from X_{kj}, where $j=1,\dots,N_{lhs}$, and $k=1,\dots,K$, rather than from S. The N_{lhs} intervals on the range of each component of X combine to form N_{lhs}^K cells that cover the sample space of X. The components of X_j are matched at random.

Figure A.1 outlines the LHS algorithm. First, one chooses the K parameters to be varied in simulation then specifies the marginal distribution of each. Let S equal this set

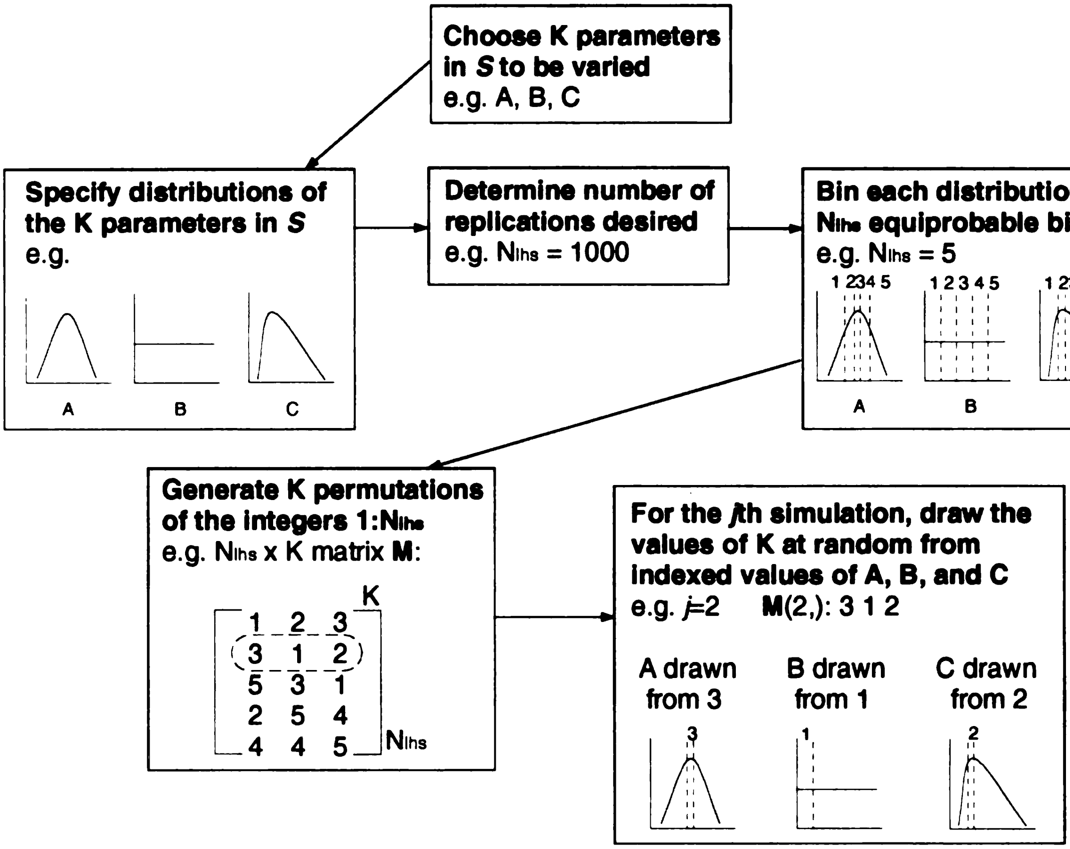
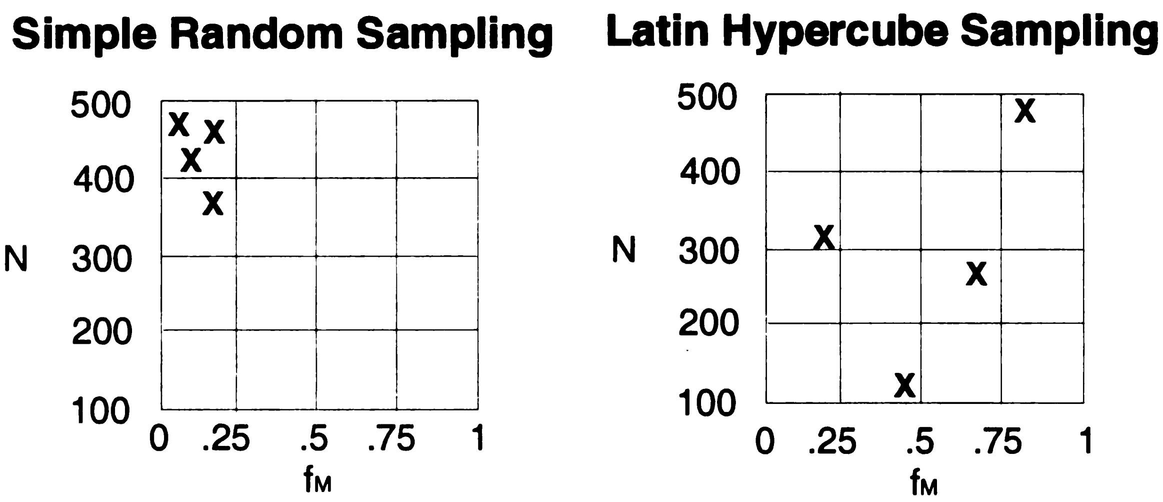

Choose K parameters in S to be varied
e.g. A, B, C
Specify distributions of the K parameters in S
e.g.
A
B
C
Determine number of replications desired
e.g. Nlhs = 1000
Bin each distribution into Nlhs equiprobable bins
e.g. Nlhs = 5
1 2 3 4 5
1 2 3 4 5
1 2 3 4 5
A
B
C
Generate K permutations of the integers 1:Nlhs
e.g. Nlhs x K matrix M:
K
1 2 3
3 1 2
5 3 1
2 5 4
4 4 5
Nlhs
For the jth simulation, draw the values of K at random from indexed values of A, B, and C
e.g. j=2 M(2,): 3 1 2
A drawn from 3
B drawn from 1
C drawn from 2
3
1
2

Figure A.1. Latin Hypercube Sampling Algorithm. Latin Hypercube Sampling involves randomly picking parameter values for each clinical trial from a subsection of each parameter's marginal distribution rather than from the entire multivariate distribution at once.

of K distributions (A, B and C in Figure A.1). Then, one specifies the number of sets of parameter values (N_{lhs}) desired. Next, each of the K marginal distributions in *S* are divided into N_{lhs} equiprobable ranges with each range indexed consecutively by the numbers 1:N_{lhs}. Then, K permutations of the integers 1:N_{lhs} are entered into a matrix (**M**) of dimension N_{lhs} × K. These indices correspond uniquely to each X_{kj} in *S*. **M**(*j*,.) (of length K) is used to determine the j^{th} set of parameter values. The value of **M**(j, k) points to the bin from which parameter k is to be drawn at random.

Simulation from randomly selected strata of the distributions in S, rather than from the entire range of the distributions in S, serves to efficiently spread points out in parameter space. For computing expectations, LHS is more efficient than Simple Random Sampling (SRS). Figure A.2 is a two-dimensional illustration of the advantage of LHS over SRS—LHS does not permit clustering of values in parameter space.


Simple Random Sampling

Latin Hypercube Sampling

Figure A.2. A Two Dimensional Illustration of the Advantage of Latin Hypercube Sampling Relative to Simple Random Sampling. Because Latin Hypercube Sampling restricts sampling of parameter space to occur only once per subsection of each parameter's marginal distribution, sampled parameter values cannot cluster in parameter space. Parameter values chosen via Simple Random Sampling are not protected in this way.

If parameters are meant to be distributed independently, it is beneficial to reduce chance correlation among the columns of **M** before using its indices to sample from *S*. Correlation between two columns, c_1 and c_2, of **M**, however small, may make it appear that a statistic that depends on c_1 depends also on c_2, when in fact, it may not. A modified Graham Schmidt Orthogonalization (GSO) procedure can reduce these unintended correlations(Owen 1994). The GSO algorithm is as follows (using Splus like pseudo code). The actual Splus subroutine is included at the end of this chapter.

```
Loop over w=1:(K-1)
      Regress M(,(w+1):K) on M(,w)
      Compute residuals
      Rank residuals
      Replace M(,(w+1):K) with ranks
Loop over x = (K-1):1
      Regress M(,x:1) on M(,(x+1))
      Compute residuals
      Rank residuals
      Replace M(,x:1) with ranks
```

One repeats both loops several times in succession until correlation between the columns of **M** fails to decrease. Graham Schmidt Orthogonalization is used for all simulations.

To further enhance parameter space coverage in this investigation, only one data set is simulated and analyzed under each parameter setting chosen. This approach spends resources covering more parameter space, rather than obtaining more precise estimates of performance at fewer points in hyperdimensional space.

Note the distinction between the use of the terms 'study' and 'trial' in this report. A 'study' is defined as an investigation of method performance over a range of parameter values. A 'trial' refers to one data set randomly simulated using a particular set of parameters from the ranges chosen in a 'study'. Thus, there are N_{lhs} 'trials' carried out in every 'study'.

Unless otherwise noted, for each study, parameter value selection, data simulation, and data analysis is replicated five hundred times by choosing N_{lhs}=100 five times (starting with different random seeds).

A.1.2 Simulation Design For Each Clinical Trial

Given θ_1, $P(C_Q|C)$, $P(C_M|C)$, and P(C), simulation of data for each trial proceeds according to the following algorithm. For i=1:N, D_{ni} is simulated by alternately assigning subjects to each possible dose. Hence, each assigned dose group is of equal size. Then, M=1 is assigned to the first N_M patients and the rest are assigned M=0. Next, C is drawn from P(C), C_Q is drawn from $P(C_Q|C)$, and C_M (for all subjects) is drawn from $P(C_M|C)$. After simulation of all compliance measures, D is computed using $D(C,D_n)$ and Y is drawn from P(Y|D). Figure A.3 illustrates the causal model for compliance and response data simulation.

Compliance is an ordinal, categorical variable with the three levels 0, .5, and 1, corresponding to zero, fifty, and one hundred percent of prescribed pills taken, respectively. When D_n consists of a placebo and a unit dose of drug, D_n:(0,1), it is referred to as the 'placebo-controlled design'. When D_n consists only of a unit dose of drug, D_n:(1), it is referred to as the 'unit dose design'. Unless otherwise noted, all reported results correspond to a placebo-controlled design.

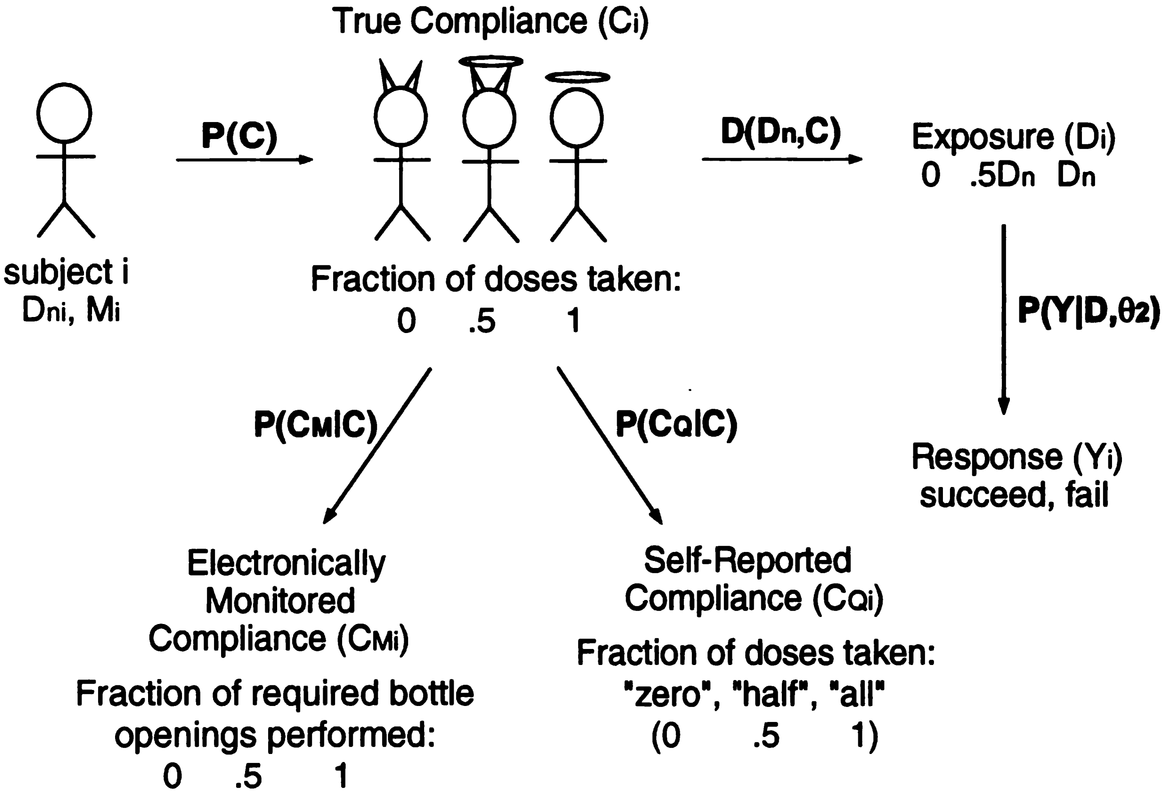


Figure A.3. Causal Model for Data Simulation. Electronically monitored compliance and self-reported compliance are each simulated conditional only on true compliance. Drug response is simulated conditional only on drug exposure.

The relationship between C, D_n and D is as follows. A patient assigned to placebo (D_n=0) has zero exposure to drug (D=0) regardless of his compliance. A patient assigned to the unit dose (D_n=1), has exposure equal to his compliance (D=C). Therefore, a subject with D=D_n and C=0 falls into the same category of exposure (D=0) as a subject assigned to placebo (formally, D(C=0,D_n)=D(C,D_n=0)=0). This is a commonly made assumption of analyses using compliance data(Efron and Feldman 1991). Given this set up, there are three possible categories of exposure, 0, .5D_n, and D_n, corresponding to zero, fifty, and one hundred percent of nominal drug exposure, respectively.

A.2 Simulation Parameters

Unless otherwise noted, P(C) assigns subjects to the three levels of C with equal probability (P(C=0)=P(C=.5)=P(C=1)=1/3). Hence, P(C) is not chosen via Latin Hypercube Sampling.

$P(C_Q|C)$ is varied in simulation studies. Figure A.4 illustrates the procedure for constructing $P(C_Q|C)$. It is formed by the linear combination of three distributions denoted $C_{Q,Q=C}$ (self-reported compliance is identical to true compliance), $C_{Q,Q\geq C}$ (self-reported compliance overestimates true compliance), and $C_{Q,Q\perp C}$ (self-reported compliance is independent of true compliance). The three distributions are weighted by the factors: $WT_{Q=C}$, $WT_{Q\geq C}$, and $WT_{Q\perp C}$. These weights must sum to 1. This constraint is satisfied by the transformation illustrated in Figure A.4 for *a*, *b*≥0, where *a* and *b* are drawn at random from uniform distributions. The value of the conditional distributions $C_{Q,Q=C}$, $C_{Q,Q\geq C}$, and $C_{Q,Q\perp C}$, are given in Table A.1. According to $C_{Q,Q\geq C}$, for example, a subject with C=.5 has a fifty percent chance of accurately self-reporting intake, or, formally $p(C_Q=.5|C=.5)=.5$. But the same individual has a fifty percent chance of saying that all of the prescribed pills were ingested $p(C_Q=1|C=.5)=.5$.

Unless otherwise noted, $P(C_M|C)$ assigns C_M as stated in **A3** under Section 4.1 of Chapter 4; C_M=C. Thus, $P(C_M|C)$ is not chosen via Latin Hypercube Sampling. To be consistent with the notation representing the components of $P(C_Q|C)$, $P(C_M|C)$ is formed by assigning a weight of 1 to the distribution $C_{M,M=C}$ (see Figure A.5).

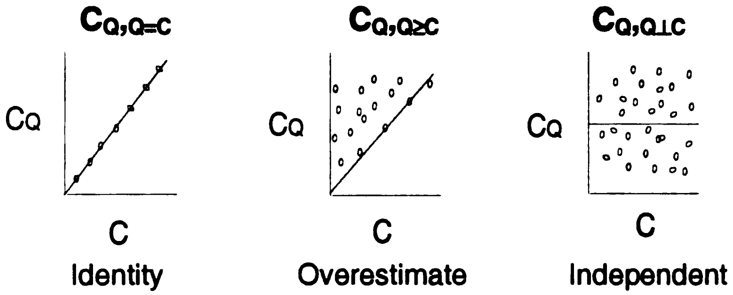


Figure A.4. Algorithm for Constructing $P(C_Q|C)$ of Simulation. The model used to simulate C_Q, $P(C_Q|C)$, varies between clinical trials. Selection of $P(C_Q|C)$ is automated by averaging three prototypical models ($C_{Q,Q=C}$, $C_{Q,Q\geq C}$, and $C_{Q,Q\perp C}$) for the relationship between C_Q and C with randomly varying weights ($WT_{Q=C}$, $WT_{Q\geq C}$, and $WT_{Q\perp C}$). $WT_{Q\perp C}$, $WT_{Q=C}$, and $WT_{Q\geq C}$ are created using the following transformation of random variables *a* and *b*:

$$WT_{Q\perp C} = a/(a+b+1),\ WT_{Q=C} = b/(a+b+1),\ WT_{Q\geq C} = 1/(a+b+1)$$

Note that *a* and *b* are drawn by Latin Hypercube Sampling from U(min(*a*),max(*a*)) and U(min(*b*),max(*b*)) such that $a,b \geq 0$.

Drug effect is a dichotomous random variable with the logit of the probability of success modeled for simulation as a linear function of exposure (D). The probability of success is arbitrarily centered on logit(.5) for the fifty percent exposure group ($D=.5D_n$). As Figure A.6 shows, specification of the drug effect model requires only one parameter—the logit of the probability of success (ρ) for another exposure group. Arbitrarily, the D=0 group was chosen to define ρ.

Table A.1. Parameter Values for the Components of $P(C_Q|C)$ of Simulation: $C_{Q,Q=C}$, $C_{Q,Q\geq C}$, and $C_{Q,Q\perp C}$. $C_{Q,Q=C}$, $C_{Q,Q\geq C}$, and $C_{Q,Q\perp C}$ are prototypical conditional distributions for C_Q given C. As an aid to interpretation of these models, under $C_{Q,Q\geq C}$, a patient who has C=0 has an equal probability of self-reporting zero, fifty percent, and perfect compliance. In contrast, under $C_{Q,Q=C}$, a patient with C=0 has a one hundred percent chance of accurately reporting zero intake.

$C_{Q,Q=C}$			C	
		0	**.5**	**1**
C_Q	**0**	1	0	0
	.5	0	1	0
	1	0	0	1

$C_{Q,Q\geq C}$			C	
		0	**.5**	**1**
C_Q	**0**	1/3	0	0
	.5	1/3	1/2	0
	1	1/3	1/2	1

$C_{Q,Q\perp C}$			C	
		0	**.5**	**1**
C_Q	**0**	1/3	1/3	1/3
	.5	1/3	1/3	1/3
	1	1/3	1/3	1/3

$C_{M,M=C}$

$$\begin{array}{ccccc} & & & \mathbf{C} & \\ & & \mathbf{0} & \mathbf{.5} & \mathbf{1} \\ \mathbf{C_M} & \mathbf{0} & 1 & 0 & 0 \\ & \mathbf{.5} & 0 & 1 & 0 \\ & \mathbf{1} & 0 & 0 & 1 \end{array}$$

Figure A.5. The Model for $P(C_M|C)$ of Simulation. Unlike $P(C_Q|C)$, $P(C_M|C)$ is fixed in simulation. C_M is generated from a model in which C_M is perfectly correlated with C. This model, referred to as $C_{M,M=C}$, is analogous to the $C_{Q,Q=C}$ component of $P(C_Q|C)$.

A.3 Performance Evaluation

Method performance is compared with respect to estimation precision—error in estimated drug effect for each trial (δ) and error in estimated drug effect for each study (Δ). Method performance is also computed in terms of power. The procedure for computing these metrics is explained below. For the most part, estimation precision and power are represented graphically. Interpretation of performance plots is explained in this section, as well.

A.3.1 Estimation Precision

Two errors are reported—δ and Δ. δ is referred to as "trial error" and Δ is referred to as "study error".

δ is computed as the absolute error on the logit scale for each method's exposure-response estimates. That is,

$\delta = |\text{logit}(\hat{\theta}_1) - \text{logit}(\theta_1)|$

The logit is simply the log odds ratio. Thus, this metric is the absolute error in the log odds ratio. Since there are three levels of exposure, there are three values of δ (δ_D, D=0, D=.5D_n, D=D_n) associated with each analysis method for each of the 500 trials in a study. Δ is the average of the 500 sets of δ for each method. Δ is reported relative to the study error for **ALL**. It is computed as the ratio of the average of $\delta_{D=0}$ and $\delta_{D=Dn}$ for each method relative to **ALL**, or formally,

$\Delta_R = \Delta^{METHOD} / \Delta^{ALL}$, where Δ=average($\delta_{D=0}$,$\delta_{D=Dn}$).

Note that $\delta_{D=.5Dn}$ is not included in the calculation of Δ_R because the true value of θ_1 for D=.5D_n, by coincidence, matches the prior penalty on θ_1 (P(Y|D=.5) = θ_{1prior} = .5). Thus, P(Y|D=.5D_n) estimation partially benefits from the prior used to stabilize estimation. Since the models are saturated, its removal is inconsequential to performance evaluation.

δ values are reported graphically as function of parameter values used in simulation. A nonparametric descriptor of the relationship between δ and the parameter—a smooth through $\delta_{D=0}$ and $\delta_{D=Dn}$ as a function of a parameter value, generically, x—is shown. To give an estimate of the precision of this estimate, a "95th percentile" confidence region is plotted around the smooth. This region is determined using the following bootstrap procedure. Pairs of error with the corresponding parameter value ($\delta_{D=0}$,$\delta_{D=Dn}$,x) are sampled with replacement to yield a data set of the same length as the original. A smooth is fit to these bootstrap samples. After repeating this procedure 500 times, the 2.5th and 97.5th percentile of the smooths (determined pointwise along x) defines the (shaded) "95th percentile" region. Note that this region does not reflect variability in the methods of analysis, but rather, the precision of the estimate of method precision as a function of the number of simulations performed.

To estimate the variability in study error (Δ_R), δ is bootstrapped 500 times for this

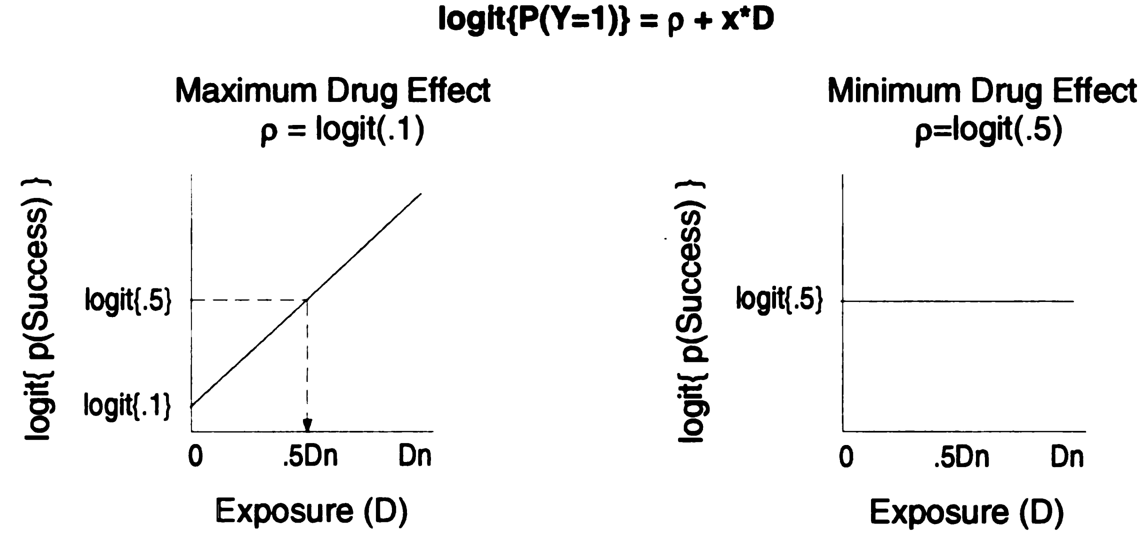


Figure A.6. Algorithm for Constructing P(Y|D) of Simulation. The model used to simulate Y, P(Y|D), varies between clinical trials. Selection of P(Y|D) is automated by assuming a log linear model for the exposure-response relationship and fixing one point in the D,Y plane around which to randomly pivot the slope of drug effect. Arbitrarily, D=.5D_n, Y=logit(.5), is the fixed point. The model is defined by choosing ρ—the value of logit{P(Y=1)} at D=0—at random by Latin Hypercube Sampling. Hence, the minimum drug effect (no drug effect) occurs when ρ=logit(.5). In the studies performed, the smallest ρ picked is .1, thus, the largest drug effect explored is P(Y|D=0)=.1 and P(Y|D=D_n)=.9.

computation. Δ_R is reported as a boxplot where the white line represents the median, the box limits represent the interquartile range, the whiskers show the most extreme values within 2.5 times the interquartile range, and outliers are indicated using horizontal lines.

A.3.2 Power

Power, or the probability of discovering that assignment to an experimental intervention causes a "significant" effect when the effect truly is "significant", commonly guides the selection of clinical trial size. Likewise, it is used as a performance metric in this investigation. Power is defined here, for each method, as the probability that a data analytic method detects a drug effect when θ_1 of simulation indicates that the drug is efficacious. Since standard power charts do not allow one to account for mismeasurement of exposure (compliance), power is determined, here, by simulation.

Formally, power is the proportion of estimates of a test statistic exceeding the critical value (δ_C) of that test statistic. To determine power through simulation, many data sets are simulated under the alternative hypothesis (H_a), analyzed under H_a and the null hypothesis (H_o), and the fraction of simulations yielding estimates of the test statistic greater than δ_C is reported. Here, the null hypothesis (H_o) is

$$P(Y|D=0) = P(Y|D=.5D_n) = P(Y|D=D_n),$$

and the alternative hypothesis (H_a) is

$$P(Y|D=0) \neq P(Y|D=.5D_n) \neq P(Y|D=D_n).$$

For simulation, of course, some actual values for the probabilities satisfying these constraints are chosen.

All of the data analysis methods maximize a likelihood, therefore, the likelihood ratio is an appropriate test statistic. The log likelihood ratio (LLR) test statistic is computed as

$$\text{LLR} = O_o - O_a$$

Where

O_o = Minimum Objective Function (-2 log likelihood) under H_o

O_a = Minimum Objective Function (-2 log likelihood) under H_a.

If LLR is 'big', the data are incompatible with H_o and H_o is rejected. LLR is asymptotically distributed chi-squared ($\text{LLR} \sim \chi^2$) and the critical value (δ_C) may be determined using a chi-squared table with

$$\chi^2_{(1-\alpha)}(q),$$

where

q = difference in number of free parameters in the H_o vs. H_a model, and

α = desired level of statistical significance.

For **ALL**, **BA**, **CD**, and **BSR**, there are three free parameters when analyzing under H_a (P(Y|D=0), P(Y|D=.5), and P(Y|D=1) are estimated) but only one free parameter when analyzing under H_o (P(Y|D=0)=P(Y|D=.5)=P(Y|D=1) is estimated). Thus, the difference, q, is 2. The result is similar for **ML**, although **ML** estimates 6 extra parameters under both H_o and H_a, representing the number of parameters required to specify $P(C|C_Q)$. In contrast, **ITT** has only two free parameters under H_a, as it only estimates P(Y|D=0) and P(Y|D=1) and the prior P(Y|D=.5) is returned. The difference, q, for **ITT** is 1. Thus, the χ^2 value with two degrees of freedom is used as the critical value for **ALL**, **BA**, **CD**, **BSR**, and **ML**, while the χ^2 value with one degree of freedom is used as the critical value for **ITT**.

Although the LLR is an appropriate test statistic, various characteristics of the data analyzed here may render the use of δ_C in published power charts inappropriate. Thus, simulation is performed to obtain customized values of δ_C, and power computed using this simulated cutoff is compared with power determined using the χ^2 critical value. The customized δ_C for each method is determined by simulating under H_o, analyzing under H_o and H_a, and computing the LLR. After repeating this procedure many times, the 95th percentile value of LLR is held as δ_C. Figure A.7 shows the algorithm used to determine power and δ_C by simulation. δ_C values are tail probabilities and require many simulations to determine accurately. The simulated value here should be considered an estimate of the true value.

The estimates of power are reported graphically as a boxplot—the white line represents the median, the box limits represent the interquartile range, the whiskers show the most extreme values within 2.5 times the interquartile range, and outliers are indicated by horizontal lines.

To determine the relationship between method power and study error (Δ_R), both power and Δ_R are computed for a set of simulations. Method performance by these two metrics is rank ordered separately and then compared.

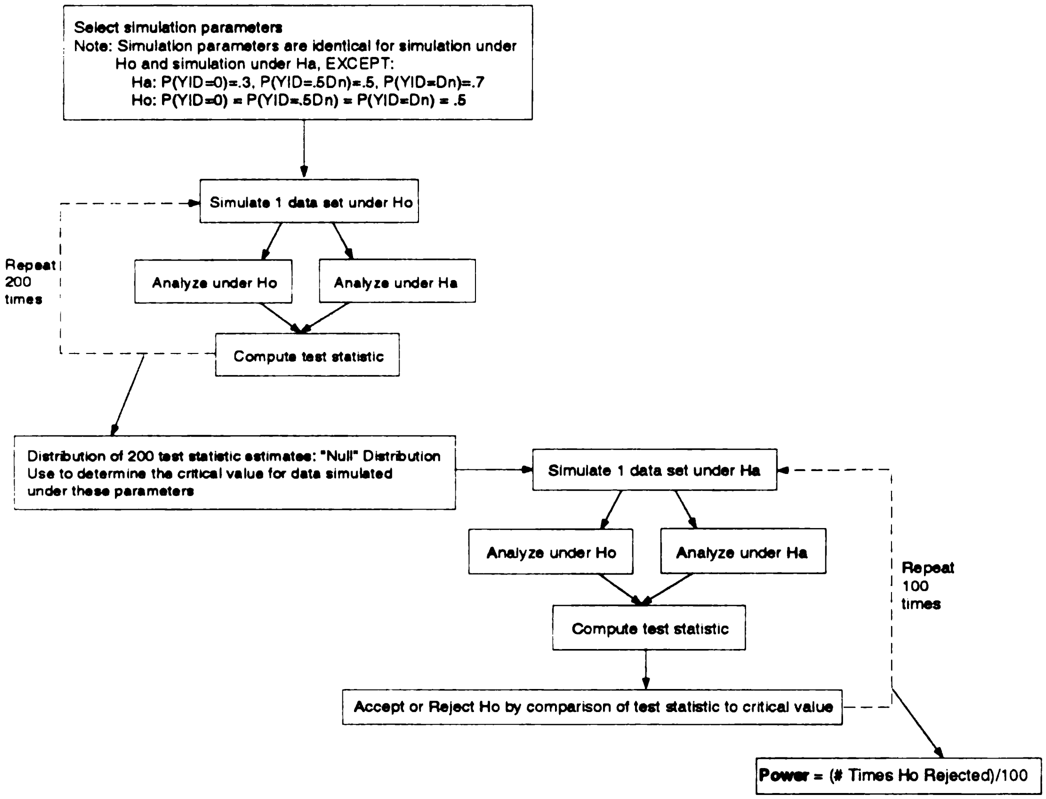

Select simulation parameters
Note: Simulation parameters are identical for simulation under Ho and simulation under Ha, EXCEPT:
Ha: P(YID=0)=.3, P(YID=.5Dn)=.5, P(YID=Dn)=.7
Ho: P(YID=0) = P(YID=.5Dn) = P(YID=Dn) = .5
Simulate 1 data set under Ho
Repeat 200 times
Analyze under Ho
Analyze under Ha
Compute test statistic
Distribution of 200 test statistic estimates: "Null" Distribution
Use to determine the critical value for data simulated under these parameters
Simulate 1 data set under Ha
Analyze under Ho
Analyze under Ha
Repeat 100 times
Compute test statistic
Accept or Reject Ho by comparison of test statistic to critical value
Power = (# Times Ho Rejected)/100

Figure A.7. Algorithm for Determining Power and the Critical Value by Simulation. Power is determined by computing the fraction of test statistics exceeding the critical value of that test statistic. The appropriateness of the χ^2 critical value is determined by comparing power determined using the χ^2 critical value to power computed using a critical value simulated under the null hypothesis. Note that:

Analysis H_a: $P(Y|D=0) \neq P(Y|D=.5D_n) \neq P(Y|D=D_n)$

Analysis H_0: $P(Y|D=0) = P(Y|D=.5D_n) = P(Y|D=D_n)$.

A.4 Splus Program for implementing Latin Hypercube Sampling with Orthogonalization to reduce correlation among parameters

The code used to generate a Latin Hypercube Sample of parameter values is given here. To generate a matrix of parameter values, simply run the function LHSwGSO in

Splus specifying the required arguments as explained in the body of the function. The GSO code is given in section A.4.2 and other necessary functions are given in section A.4.3. All of the Splus code listed in this appendix must be input into Splus for the LHSwGSO function to work.

A.4.1 LHS code

```
LHSwGSO _ function(Nlhs, breakpoints, outfile, noGSO) {


### This function is used to generate an Nlhs x k matrix of
### parameter values where
###   k = Number of parameters to be sampled
###   Nlhs = Number of times to sample parameters
### The seed for generating random numbers must be set external to
### the function.
###
### The parameter values are sampled by Latin Hypercube Sampling:
### a pseudo-random multidimensional stratified sampling approach.
### For more information refer to:
###   Iman, et. al. Risk Analysis, 1988, 8(1):71-90.
###   Loh, W.L. Annals of Statistics, Oct, 1996, 24(5): 2058-2080.
###   McKay, et. al. Technometrics, Feb 2000, 42(1):55-61.
###   McKay, et. al. Technometrics, May 1979, 21(2):239-245.
###
### An Orthogonalization procedure is used to reduce correlation
### among Latin Hypercube sampled parameters.
### For more information refer to:
###   Owen, A.B. Journal of the American Statistical Association
```

```
      Dec 1994, 89(428):1517-1522.
#
# Arguments of the function:
#
# Nlhs
#     = Number of times to sample parameters
#           Notes:
#           1. Nlhs must be greater than the number of
#              parameters to be sampled
#           2. If Nlhs < 10, Graham Schmidt Orthogonalization
#              is not used
#
# breakpoints
#     = List of parameters and their corresponding breakpoints to be
#       used in stratified sampling
#           Breakpoints are determined by dividing each parameter's
#           distribution into Nlhs equiprobable ranges.
#           Length of each parameter's breakpoints  = Nlhs+1.
#           Note:
#           1. Must write parameter breakpoints in ascending order
#           2. Breakpoints for a given parameter must be unique.
#           3. The length of each vector of breakpoints must be
#              identical for all parameters.
#           4. The name of the parameter is taken as the name given
#              to each element of the list.
#
#           e.g. breakpoints = list(CL=c(.2,.4,.6,.8),
#                 V=c(100,200,300,400))
#           In this example,
#                 There are 2 parameters: CL, V
#                 3 sampled values are desired (Nlhs=3)
#
#           *** There is one default sampling distribution:
#                 uniform sampling
#           If it is desired to have parameter values sampled
#           uniformly between some minimum and maximum value,
#           specify the minimum and maximum value, and set Nlhs
#           to the desired number of samples.
#           e.g. Since the above distributions are uniform, could
#           also have specified:
#                 Nlhs=3,
#                 breakpoints = list(CL=c(.2,.8), V=c(100,400))
#           This approach only works if all parameters are to be
#           sampled from a uniform distribution.
#
# outfile
#     = Name of output file
#           e.g. outfile=c("filename")
# noGSO
#     = Logical variable: should GSO be omitted?
#           If noGSO=T, LHS is performed without orthogonalization

# Initialization
   kpars _ length(breakpoints)
   parammat _ matrix(NA, nrow=Nlhs, ncol=kpars)

# Checking arguments of the function:
   # Test 1:
```

```
lenbp _ rep(NA,kpars)
lenuniqbp _ rep(NA,kpars)

for (i in 1:kpars) {
   lenbp[i] _ length(breakpoints[[i]])
}

# A) Are all breakpoint vectors the same length?
uniqlenbp _ unique(lenbp)
if (length(uniqlenbp) != 1) {
   write("Specified breakpoint vectors are of unequal length",
       file="LHSerror")
   stop("Specified breakpoint vectors are of unequal length")
}
# B) Are there 2 or more numbers in breakpoints?
if (uniqlenbp < 2) {
   write("May not have fewer than 2 breakpoints", file="LHSerror")
   stop("May not have fewer than 2 breakpoints")
}

# C) Does the length of breakpoints appropriately correspond with
#    Nlhs?
if (uniqlenbp > 2) {
   if (uniqlenbp != Nlhs+1) {
      write("Number of breakpoints specified must be Nlhs+1",
         file="LHSerror")
      stop("Number of breakpoints specified must be Nlhs+1")
   }
}
```

```
   # Test 2: Are all breakpoints unique?
   for (i in 1:kpars) {
      lenuniqbp[i] _ length(unique(breakpoints[[i]]))
   }
   uniqlenuniqbp _ unique(lenuniqbp)

   if (length(uniqlenuniqbp) != 1) {
      write("At least one specified breakpoint vector has a nonunique
value",
          file="LHSerror")
      stop("At least one specified breakpoint vector has a nonunique
value")
   }

   # Test 3: Are breakpoints listed in ascending order?
   for (i in 1:kpars) {
      orderbp _ order(breakpoints[[i]])
      if (uniqlenbp>2) {
       testorder _ orderbp != 1:(Nlhs+1)
         uniqtestorder _ unique(testorder)

         if ( length(uniqtestorder) != 1) {
          write("Breakpoints not listed in ascending order",
                  file="LHSerror")
          stop("Breakpoints not listed in ascending order")
       }
      }
   }
   # Test 4: Was Nlhs specified > number of parameters to be sampled?
   if (Nlhs <= kpars) {
      write("Nlhs must be > number of parameters specified",
            file="LHSerror")
      stop("Nlhs must be > number of parameters specified")
   }

# Action
   # Create a look up table for each parameter
   lutlist _ vector("list", kpars)
   names(lutlist) _ names(breakpoints)
   if (uniqlenbp==2) {
      for (i in 1:kpars) {
         lutlist[[i]] _ makelut(breakpoints[[i]][1],
                        breakpoints[[i]][2], Nlhs)
      }
   }
   else {
      for (i in 1:kpars) {
         minbp _ min(breakpoints[[i]])
         maxbp _ max(breakpoints[[i]])

         bp.tmp1 _ breakpoints[[i]][-1]
         bp.tmp2 _ breakpoints[[i]][-uniqlenbp]

         bp.tmp _ intersect(bp.tmp1, bp.tmp2)
         bp.tmp _ matrix(c(bp.tmp,bp.tmp), ncol=2)
         bp.tmp _ matrix(c(minbp, bp.tmp, maxbp), byrow=T, nrow=2)
      lutlist[[i]] _ bp.tmp
   }
}

# Create an Nlhsxk matrix of random permutations
# If Nlhs > 10, they are adjusted by Graham Schmidt
# Orthogonalization

matNlhsxk _ matrix(NA, nrow=Nlhs, ncol=kpars)

test1ind _ Nlhs > 10
test2ind _ noGSO == F
test12ind _ test1ind*test2ind

if (test12ind == 1) {
   matNlhsxk _ decrLHScor(Nlhs, kpars, .0000000001, 12)
}
else {
   for (i in 1:kpars) {
      matNlhsxk[,i] _ randindx(Nlhs)
   }
}
# Simulate random numbers between limits of parameters
for (i in 1:Nlhs) {
   for (j in 1:kpars) {
      binindex _ matNlhsxk[i,j]

      if (lutlist[[j]][1,binindex] < lutlist[[j]][2,binindex]) {
            parammat[i,j] _ runif(1, lutlist[[j]][1,binindex],
                                lutlist[[j]][2,binindex])
         }
         else {
            parammat[i,j] _ runif(1, lutlist[[j]][2,binindex],
                                lutlist[[j]][1,binindex])
         }
      }
   }

# Output
   # Write parammat to file
   write(names(lutlist), file=outfile, ncolumns=kpars)
   write(t(parammat), file=outfile, ncolumns=kpars, append=T)
   # Return Splus matrix
parammat
}
```

A.4.2 GSO code

```
decrLHScor _ function(Nlhs, numparams, minRMS, maxcountpass) {

### This function carries out an Orthogonalization procedure
### (Owen, A.B., JASA(1994)89:1517-1522) to reduce correlations
### between columns of permutations created for Latin Hypercube
### Sampling.

### Graham Schmidt Orthogonalization
###   Given:
```

```
###         p = # parameters     (p = numparams)
###         P > p
###         n = # studies to run (n = Nlhs)

###   a) Make an n x P matrix of n*P draws from a uniform
###         distribution
###   b) for (i in 2:P) {
###         Regress nxPmatrix[,i] on nxPmatrix[,1:i-1]
###         ***Since residuals are uncorrelated with x, turn
###            the residuals into an entry in the matrix.
###         Rank the correlations.
###         Replace nxPmatrix[,i] with ranks
###      }
###      for (i in (P-1):1) {
###         Regress nxPmatrix[,i] on nxPmatrix[,P:i]
###         ***Since the residuals are uncorrelated with x, turn
###               the residuals into an entry in the matrix.
###         Rank the correlations.
###         Replace nxPmatrix[,i] with ranks
###      }
###         c) Check root mean squared correl.
###         d) Repeat until a specified number of iterations are
###               completed (maxcountpass) or the RMS stops
###               decreasing a specified amount (RMSdelmin).

# Arguments of the function:
# Nlhs
#           = Number of times to sample by LHS
#                 Note: Nlhs must be > numparams
```

```
#
# numparams
#            = Number of parameters to be sampled
#
# minRMS
#            = Minimum change in RMS to be considered nonconverged
#
# maxcountpass
#            = Maximum number of passes for orthogonalization
#
   # Test Input:
   if (Nlhs <= numparams) {
      write("Nlhs must be > numparams", file="LHSerror", append=T)
      stop("Nlhs must be > numparams")
   }

   # Initialization:
   # Create a matrix of Nlhs*Pnumparams draws from a uniform
   # distribution. (Empirical evidence shows that orthogonalization
   # carried out on ncol=Pnumparams where Pnumparams > numparams
   # yields less correlation than orthogonalization carried out on
   # ncol=numparams. After orthogonalization, a matrix with
   # ncol=numparams is returned by randomly selecting numparams
   # columns from ncol=Pnumparams.)

   Pnumparams _ round(numparams*1.5)
   matnxP _ matrix(runif(Nlhs*Pnumparams), nrow=Nlhs)

   # Initialize counter for number of passes and evaluating RMS
```

```
RMS _ 100
countpass _ 1

# Action:
# Loop back and forth over columns of matnxP regressing, taking
# residuals, computing residuals, reassigning ranks

while (RMS > minRMS && countpass <= maxcountpass) {

   countpass _ countpass + 1

   # Calculate RMS
   RMSdenom _ (Pnumparams-1)*Pnumparams/2
   SQcor _ matrix(0, nrow=Pnumparams, ncol=Pnumparams)

   for (ij in 2:Pnumparams) {
      for (kl in 1:(ij-1)) {
         SQcor[ij,kl] _ (cor(matnxP[,ij],matnxP[,kl]))^2
      }
   }

   RMS _ (sum(SQcor)/RMSdenom)^.5
   print(paste("RMS = ", RMS))

   for (wx in 2:Pnumparams) {
      for (yz in 1:(wx-1)) {
       matnxP[,yz] _ rank(lsfit(matnxP[,wx],
                     matnxP[,yz])$residuals)
      }
   }

   for (wx in (Pnumparams-1):1) {
      for (yz in Pnumparams:(wx+1)) {
       matnxP[,yz] _ rank(lsfit(matnxP[,wx],
                     matnxP[,yz])$residuals)
      }
   }

   # Calculate RMS
   RMSdenom _ (Pnumparams-1)*Pnumparams/2
   SQcor _ matrix(0, nrow=Pnumparams, ncol=Pnumparams)

   for (ij in 2:Pnumparams) {
      for (kl in 1:(ij-1)) {
         SQcor[ij,kl] _ (cor(matnxP[,ij],matnxP[,kl]))^2
      }
   }

   RMS _ (sum(SQcor)/RMSdenom)^.5

   print(paste("RMS = ", RMS))
}

# Randomly select p columns
pickcols _ order(runif(Pnumparams))
pickcols _ pickcols[1:numparams]

matnxp _ matnxP[,pickcols]
```

```
   if (numparams==1) {
      matnxp _ matrix(matnxp, ncol=1)
   }
matnxp
}
```

A.4.3 Other Necessary Functions

```
makelut _ function(mindist,maxdist,Nlhs) {

### Tis function creates a look up table for values of a uniform
### distribution corresponding to an Nlhs index

# mindist   Minimum value of uniform distribution to stratify
# maxdist   Maximum value of uniform distribution to stratify
# Nlhs              Number of Latin Hypercube strata

   lut _ matrix(NA,nrow=2,ncol=Nlhs)
   boundaries _ rep(NA,Nlhs+1)
   increments _ (maxdist - mindist)/Nlhs
   for (i in 0:Nlhs) {
      boundaries[i+1] _ mindist+(increments*i)
   }

   for (i in 1:Nlhs) {
      lut[,i] _  c(boundaries[i],boundaries[i+1])
   }
lut
}

randindx _ function(Nlhs) {

### This function generates indices for LHS sampling in random order

# Argument of the function:
# Nlhs            Number of strata for LHS sampling

   tmp _ runif(Nlhs)
   perm _ match(tmp,sort(tmp))

   # The following just safeguards against having an index
   # repeated twice
   while (length(unique(perm))!=Nlhs) {
      tmp _ runif(Nlhs)
      perm _ match(tmp,sort(tmp))
   }
perm
}
```

A.5 Technical Details

All data simulation, analyses, and graphic displays are produced using Splus version 5.1 Release 1 distributed by MathSoft, Inc.. The computations are performed on a Sun system, Solaris 1. The search algorithm used is the nlmin() function in Splus. $P(C|C_Q)$ is parameterized in terms of $x^2/(1+x^2)$ and $P(Y|D)$ is estimated on the logit scale.

A.6 References

Efron, B. and D. Feldman (1991). "Compliance as an Explanatory Variable in Clinical Trials." Journal of the American Statistical Association **86**: 9-22.

Iman, R. L. and J. C. Helton (1988). "An Investigation of Uncertainty and Sensitivity Analysis Techniques for Computer Models." Risk Analysis **8**(1): 71-90.

McKay, M., R. Beckman and W. Conover (2000). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code." Technometrics **V42**(N1): 55-61.

McKay, M. D., R. J. Beckman and W. J. Conover (1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code." Technometrics **21**(2): 239-245.

Owen, A. B. (1994). "Controlling Correlations in Latin Hypercube Samples." Journal of the American Statistical Association **89**(428): 1517-1522.