Extensions of the Dynamic Propensity Score Adjustment for Longitudinal Data

BY

COLIN C. HUBBARD
B.A., Carleton College, 1985
M.A., Loyola University Chicago, 1996
M.S., University of Illinois at Chicago, 2002

THESIS

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Public Health Sciences in the Graduate College of the University of Illinois at Chicago, 2017

Chicago, Illinois

Defense Committee:
Hakan Demirtas, Chair and Advisor
Donald Hedeker, University of Chicago
Sally Freels
Li Liu
Stanley Sclove, Information & Decision Sciences
ACKNOWLEDGMENTS

I gratefully acknowledge the guidance, support, and encouragement of Donald Hedeker and Hakan Demirtas, who, in seriatim, served as the Chair of my committee and principal advisor. I am indebted to you both for your patient guidance and your prudent use of the occasional, well-timed nudge to keep me on track. I would also like to acknowledge the other members of my committee, Sally Freels, Li Liu, and Stanley Sclove, for their service and support. My wonderful parents, Jacquelyn Hubbard and the late Colin P. Hubbard, M.D., have always been a source of love, support, and inspiration, and my debt to you both is immeasurable but gratefully acknowledged. My children, Elias and Bella, have been staunch supporters of my efforts to complete this dissertation, a project that is older than Bella and not much younger than Eli. Neither has a memory of a father who was not in school and didn’t have a scholastic deadline to meet (or miss). I love you both ... to the moon and back! Lastly, I am most indebted to my wife, Rachel Rosner, who has patiently supported and encouraged me throughout; who has endured many more than a hundred indecisions, visions, and revisions; who has sacrificed so much so that I could continue to labor on this project. As I have said before, getting married to you was the best decision you ever made for me. I mean it, though.

It goes without saying, but I will say it anyway, that any mistakes herein are entirely my own.

CCH
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
</tr>
<tr>
<td>2</td>
<td>LITERATURE REVIEW</td>
</tr>
<tr>
<td>2.1</td>
<td>Propensity Scores</td>
</tr>
<tr>
<td>2.2</td>
<td>Dynamic Adaptation of the Propensity Score</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Mixed-effects Ordinal Regression Model for Propensity for Treatment</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Stratified Treatment Effectiveness Analysis</td>
</tr>
<tr>
<td>2.3</td>
<td>Consequences of Misspecification of the Propensity Score</td>
</tr>
<tr>
<td>2.4</td>
<td>Robust Alternatives to the Normal Model</td>
</tr>
<tr>
<td>3</td>
<td>PROPOSED MODEL</td>
</tr>
<tr>
<td>3.0.1</td>
<td>Mixed-effects Ordinal Regression Model for Propensity for Treatment</td>
</tr>
<tr>
<td>3.0.2</td>
<td>Propensity for Treatment Model - The Heterogeneity Model</td>
</tr>
<tr>
<td>3.0.3</td>
<td>Propensity for Treatment Model - Estimation and Inference</td>
</tr>
<tr>
<td>3.0.4</td>
<td>Propensity for Treatment Model - Empirical Bayes Estimation</td>
</tr>
<tr>
<td>3.0.5</td>
<td>Propensity for Treatment Model - Ranking the Propensity for Treatment</td>
</tr>
<tr>
<td>3.0.6</td>
<td>Stratified Linear Mixed-Effects Model for Effectiveness</td>
</tr>
<tr>
<td>4</td>
<td>SIMULATION STUDY</td>
</tr>
<tr>
<td>4.1</td>
<td>Overview</td>
</tr>
<tr>
<td>4.2</td>
<td>Simulation Specifications</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Distribution of Random Effects</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Model for Propensity for Treatment ( (M_{Treat}) )</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Model for Treatment Effectiveness ( (M_{Resp}) )</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Simulation Size</td>
</tr>
<tr>
<td>4.3</td>
<td>Evaluation of Model Performance</td>
</tr>
<tr>
<td>4.4</td>
<td>Simulation Study Results</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Normal Random Effects</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Non-normal Random Effects</td>
</tr>
<tr>
<td>4.5</td>
<td>Is the Ordinal Mixed Effects Model Robust to Violations of the Normality Assumption?</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Simulation Design</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Evaluation of Performance</td>
</tr>
<tr>
<td>4.5.3</td>
<td>Simulation Results</td>
</tr>
<tr>
<td>5</td>
<td>ALTERNATIVES TO STRATIFICATION IN THE USE OF THE ORDINAL DYNAMIC PROPENSITY SCORE</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>5.2</td>
<td>Alternatives to Subclassification on the Propensity Score</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS (Continued)

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2.1 Linear Mixed Effects Regression</td>
<td>49</td>
</tr>
<tr>
<td>5.2.2 Inverse Probability Weighting</td>
<td>50</td>
</tr>
<tr>
<td>5.2.3 Matching on the Propensity Score</td>
<td>51</td>
</tr>
<tr>
<td>5.2.4 Covariate Adjustment using the Propensity Score</td>
<td>53</td>
</tr>
<tr>
<td>5.3 Simulation Design</td>
<td>54</td>
</tr>
<tr>
<td>5.4 Simulation Results</td>
<td>55</td>
</tr>
<tr>
<td>6 ANALYSIS OF NIMH COLLABORATIVE DEPRESSION STUDY</td>
<td>59</td>
</tr>
<tr>
<td>6.1 The CDS Data</td>
<td>59</td>
</tr>
<tr>
<td>6.2 Results</td>
<td>62</td>
</tr>
<tr>
<td>6.2.1 Model for Propensity for Treatment Intensity</td>
<td>62</td>
</tr>
<tr>
<td>6.2.2 Evaluation of Propensity-Adjusted Balance</td>
<td>63</td>
</tr>
<tr>
<td>6.2.3 Model for Treatment Effectiveness</td>
<td>65</td>
</tr>
<tr>
<td>7 CONCLUSION</td>
<td>69</td>
</tr>
<tr>
<td>CITED LITERATURE</td>
<td>72</td>
</tr>
<tr>
<td>VITA</td>
<td>79</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>37</td>
</tr>
<tr>
<td>II</td>
<td>38</td>
</tr>
<tr>
<td>III</td>
<td>39</td>
</tr>
<tr>
<td>IV</td>
<td>44</td>
</tr>
<tr>
<td>V</td>
<td>45</td>
</tr>
<tr>
<td>VI</td>
<td>46</td>
</tr>
<tr>
<td>VII</td>
<td>56</td>
</tr>
<tr>
<td>VIII</td>
<td>58</td>
</tr>
<tr>
<td>IX</td>
<td>61</td>
</tr>
<tr>
<td>X</td>
<td>67</td>
</tr>
<tr>
<td>XI</td>
<td>67</td>
</tr>
<tr>
<td>XII</td>
<td>68</td>
</tr>
<tr>
<td>XIII</td>
<td>68</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Description</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mixture of Two Normals</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>Log Normal</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>Mean Symptom Severity by Treatment Intensity</td>
<td>62</td>
</tr>
<tr>
<td>4</td>
<td>Box Plot of Duration by Treatment Intensity</td>
<td>63</td>
</tr>
<tr>
<td>5</td>
<td>Trajectory by Treatment Intensity</td>
<td>64</td>
</tr>
<tr>
<td>6</td>
<td>Age by Treatment Intensity</td>
<td>65</td>
</tr>
<tr>
<td>7</td>
<td>Site by Treatment Intensity</td>
<td>66</td>
</tr>
</tbody>
</table>
SUMMARY

Dynamic propensity adjustment models are extensions of the propensity score adjustment strategy to longitudinal observational study data, in which observations on the same subjects are gathered repeatedly over time. If the treatment or exposure is ordinal, random intercept ordinal regression models are used to model the propensity for ordinal treatment with measured covariates, both baseline and time-varying, as predictors in the model. Level-1 observations are stratified based on quintiles of their estimated dynamic propensity score, and mixed effects regression models are fit within each propensity stratum, with treatment and measured covariates as predictors for the outcome of interest. In the absence of significant propensity quintile-by-treatment interaction, a common treatment effect is estimated using a weighted, pooled estimator.

In this dissertation we describe several extensions of the dynamic propensity score methodology, including a method which employs a model that is robust to misspecification of the random effects distribution. We develop this model and analyze its performance using a Monte Carlo simulation study. We find that the dynamic propensity adjustment is robust to violations of random effect distributional assumptions, and its performance is not improved by using more flexible, but computationally demanding methods of modeling the dynamic propensity score. In an effort to understand these results, we conduct a simulation study to assess the degree to which ordinal random effects models are robust to misspecification of the random effects distribution. In addition, we develop methods that use the dynamic propensity score as an index for matching, weighting, and direct adjustment and evaluate each methods performance using simulation studies. We find that these methods perform well when the effect size of confounders is modest, but no better than the adjustment method based on stratification. When the effect size
is large, stratification outperforms these methods. Finally, we analyze data from The National Institute of Mental Health Collaborative Depression Study, a longitudinal observational study of the treatment of major affective disorders, using stratification on the estimated dynamic propensity score.
CHAPTER 1

INTRODUCTION

The dynamic propensity adjustment models were developed by Leon and Hedeker in a series of articles published over the last decade (Leon et al., 2001; Leon and Hedeker, 2005; Leon et al., 2007; Leon and Hedeker, 2007a; Leon and Hedeker, 2007b; Leon and Hedeker, 2011; Leon et al., 2012). The dynamic propensity adjustment is an extension of the propensity score methodology to longitudinal data settings in which the treatment is partitioned into ordinal doses (e.g., “low”, “medium”, and “high”). A random intercept ordinal regression model which assumes normality of the random effects is fit with the ordinal dose regressed on the measured covariates, both baseline and time-varying. Using the empirical Bayes estimates of the random effects and the estimated regression coefficients, an estimated propensity for treatment intensity is computed for each measurement occasion for each subject. Observations (not subjects) are then stratified by propensity quintiles, and then propensity stratum-specific random effects models are fit with the effectiveness measure regressed on the dose. (The latter models can be either linear mixed models or generalized linear mixed models, depending on the nature of the outcome measure.) Finally, the treatment effectiveness parameter estimates from each quintile are pooled, to obtain an overall estimate of treatment effectiveness.

The context in which these models were developed is an analysis of long-term, observational study of the effectiveness of somatic antidepressant treatment for reducing the severity of depressive symptoms in patients with major depressive disorder.

Because the propensity model for treatment intensity assumes normality of the random effects, it is subject to the same potential biases as most generalized linear mixed models that
make the same assumption. There appears to be considerable, unresolved controversy in the literature regarding the impact of random effects misspecification on estimation of the fixed parameters and the empirical Bayes estimates of the random effects, both of which are used to estimate the predicted propensity for treatment in the dynamic propensity approach. There have been several simulation studies that have reported that estimation of these parameters can be sensitive to misspecification of the random effects distribution. In addition, there is no specific simulation study which examines the impact of the misspecification of the random effects distribution on estimation in mixed effects ordinal regression models.

In this thesis we describe an adaptation of the dynamic propensity model approach that employs mixed models that use a more flexible approach to modeling the random effects distribution. More specifically, the “heterogeneity model” (Verbeke and Lesaffre, 1996), developed originally in the context of linear mixed models, is adapted for use in Leon and Hedeker’s dynamic adjustment propensity model.

This thesis is organized as follows:

In Chapter 2 a literature review is presented which first describes the development of the concept of the propensity score. Second, we review Leon and Hedeker’s dynamic adaptation of the propensity score for longitudinal data. Third, we review the statistical literature regarding the consequences of misspecification of the random effects distribution in mixed models on estimation and prediction. Finally, we describe various solutions that have been developed to the problem of random effects misspecification in mixed models.

In Chapter 3, we describe the model which incorporates the heterogeneity model into the dynamic adaptation of the propensity score.

In Chapter 4, we describe the set up and results of several Monte Carlo simulations studies we conducted to examine the performance of the proposed model relative to the existing model
and to assess the robustness of ordinal mixed effects models, in general, to violations of its
distributional assumptions.

In Chapter 5, we describe alternative uses of the dynamic propensity score for ordinal
treatments other than stratification. We develop weighting, matching, and covariate adjustment
strategies that use the dynamic propensity score. We also conduct a simulation study to assess
their relative performance.

In Chapter 6, we analyze data from The National Institute of Mental Health (NIMH)
Collaborative Depression Study (CDS), a longitudinal observational study of the treatment of
major affective disorders.

In Chapter 7, we summarize our findings and make some suggestions for further research.
CHAPTER 2

LITERATURE REVIEW

2.1 Propensity Scores

In contrast to a randomized experiment, in an observational study the assignment of treatments to study subjects is not under the control of the experimenter. Nonetheless, observational studies may be undertaken for a variety of reasons, including cost limitations, ethical prohibitions against assignment of harmful treatment or withholding of known efficacious treatment, or because data has already been collected and is available for immediate analysis, whereas the results of a randomized trial might take years to become available (Rubin, 1986). Furthermore, in certain contexts, observational studies may be preferred, or at least seen as complementary to, randomized studies. For instance, in the context of comparative effectiveness research, in which treatment options are compared in "real world" settings, the results from observational studies may be considered a better measure of treatment effectiveness in the target population than the evidence from randomized trials (Armstrong, 2012; Flay, 1986). Such a situation can arise, for example, in a randomized clinical trial in which the exclusion criteria strictly limits the types of individuals eligible to participate in the study. In these circumstances, the subjects of an observational trial may be more representative of the population at large than those of a randomized trial.

Of course, the advantages of a randomized trial are well-known. Because treatment assignment is under the control of the experimenter, who usually employs a random mechanism to assign treatment, randomized treatment assignment tends, at least in large experiments, to produce similar treatment groups, with fairly evenly distributed covariates, both measured and
unmeasured. Because treatment groups tend to be comparable in a randomized study, and therefore the distribution of independent risk factors for the outcome of interest is balanced between treatment groups, confounding is far less of a concern than in a non-randomized study.

To illustrate this point further, we will employ the so-called potential outcomes model of causation (Holland, 1986; Little and Rubin, 2000; Rubin, 1986; Morgan and Winship, 2007). Let us imagine a study with two treatments, where $Z = 1$ if a subject is assigned to treatment and $Z = 0$ if assigned to control. In the potential outcomes framework, each subject is assumed to have — irrespective of the treatment he or she actually receives — a fixed potential, possibly counterfactual, outcome for each possible treatment assignment. In the case of a binary treatment, each subject is presumed to have two potential outcomes, $R^1$, the response if the subject receives the experimental treatment, and $R^0$, the subject’s response were he or she to receive the control treatment. Although each subject has two potential outcomes, only one is observed for each subject, since a subject cannot be assigned to both treatments (unless it is a crossover design). Indeed, what we actually observe is the outcome conditional on treatment assignment, i.e., we either observe $(R^1|Z = 1)$ or $(R^0|Z = 0)$; $(R^1|Z = 0)$ and $(R^0|Z = 1)$ are not observable. It is usually assumed that the potential outcomes of one subject are independent of the treatment status of different subjects (no interference between subjects) and are unaffected by the manner in which treatment is allocated. The latter assumption is known as the stable unit treatment value assumption (Rubin, 1980).

In a randomized study it is reasonable to assume that due to the process of random assignment, the treatment assignment, $Z$, is independent of the potential outcomes or responses given observed and unobserved covariates. Randomized trials therefore enjoy the property of what Rosenbaum and Rubin have termed strong ignorability (Rosenbaum, 1984a). A treatment assignment is strongly ignorable given observed covariates if (a) the potential responses are con-
ditionally independent of treatment assignment given the observed covariates, and (b) at each combination of observed covariates there is a positive probability of receiving each treatment. Written formally, treatment assignment is strongly ignorable given a vector of covariates $X$, if

$$p(Z|R^t, X) = p(Z|X)$$

and,

$$0 < Pr(Z = t) < 1,$$

for $t = 0, 1$ and for all $X$.

Propensity score methodology was introduced by Rosenbaum and Rubin (1983) to deal with the problem of non-comparable treatment groups in observational studies. Rosenbaum and Rubin defined the propensity score as the conditional probability of assignment to an experimental treatment given a vector of observed covariates. Suppose that $N$ subjects are included in an observational study, and each subject is assigned to one of two treatments (keeping in mind that in an observational study the "assignment" of treatment or exposure is outside the control of the investigator). In addition, prior to treatment a vector of covariates, $X$, is observed for each subject. As above, let $Z = 1$ if a particular subject is assigned to experimental treatment, and let $Z = 0$ if assigned to a control or reference treatment. The propensity score $e(X)$ for a subject with covariate vector $X$ is:

$$e(X) = Pr(Z = 1|X).$$

While $X$ is a $p$-dimensional vector, $e(X)$ is a scalar. The propensity score acts as a balancing score for $X$, in that the conditional distribution of the covariates used to construct the score,
\( X \), given \( e(X) \), is the same for treated \((Z = 1)\) and control groups \((Z = 0)\). If treatment assignment is strongly ignorable given the observed covariates, then it is strongly ignorable given the propensity score \( e(X) \) (Rosenbaum and Rubin, 1983).

In a randomized experiment the propensity score (i.e., the probability of being assigned to the treatment group) is known, but in an observational study the propensity score of each subject must be estimated from the data, using a statistical model, such as logistic regression. Theoretical work and simulation have shown that estimated propensity scores are actually better at removing bias than true propensity scores, because adjustment for the estimated propensity score removes both systematic bias from covariate imbalance (bias due to confounding) and imbalance due to chance alone (sampling error) (Rosenbaum and Joffe, 1999). In contrast, adjustment with the true propensity score only removes systematic bias.

If treatment assignment is strongly ignorable, Rosenbaum and Rubin have shown that forms of adjustment, such as matching (Rosenbaum and Rubin, 1985), subclassification (Rosenbaum and Rubin, 1984), and covariance adjustment (Rosenbaum, 1987) on the propensity score produce unbiased estimates of treatment effects (Rosenbaum and Rubin, 1983). The advantage of the propensity score is that it is much easier to adjust for a scalar function of the covariates than it is to adjust for a vector of covariates. As Rosenbaum and Rubin (1983) point out, the number of subclasses created by subclassification on a vector of covariates increases exponentially with the dimension of the covariate vector. In addition, as the number of covariates grows, the problem of sparse subclasses increases. Following Cochran (1968), Rosenbaum and Rubin (1984b) argue that subclassification at the quintiles of the distribution of the estimated propensity score is adequate to obtain substantial improvement in the balance of observed covariates, although the adequacy of the covariate balance achieved can and should be examined in each case.
2.2 Dynamic Adaptation of the Propensity Score

Rosenbaum and Rubin’s propensity score methodology was developed for a dichotomous, time-invariant treatment. In a series of articles, Leon and colleagues (Leon et al., 2001; Leon and Hedeker, 2005; Leon et al., 2007; Leon and Hedeker, 2007a; Leon and Hedeker, 2007b; Leon et al., 2012) have adapted and extended the propensity score methodology for use in the analysis of longitudinally collected data in which the treatment of interest varies over time and is ordinal in nature. In a longitudinal study, repeated observations of the same variables are made on subjects over time. On each measurement occasion covariate values are recorded, as are the treatment level and the outcome of interest, which are typically time-varying. For each subject $i$, $(i = 1, \ldots, N)$ in a longitudinal study, there is a $n_i \times p$ matrix of covariates, $X_i$, in which the $i$-th row represents the vector of covariates measured on occasion $j$, $j = 1, \ldots, n_i$; a vector of treatments, $T_{ij}$; and a vector of outcomes, $Y_{ij}$.

Leon et al. (2001), followed by Leon and Hedeker (2005), introduced several innovations in their adaptation of the propensity score methodology. First, they defined a propensity score for ordinal categorical treatments, such as none, small dose, moderate dose, and maximum dose. Second, they proposed a two-stage data analytic strategy in which (a) the propensity for treatment intensity is modeled using a mixed effects ordinal regression model, and then (b) the effectiveness of ordered categorical treatments is evaluated using a stratified analysis based upon the mixed-effects ordinal propensity score.

2.2.1 Mixed-effects Ordinal Regression Model for Propensity for Treatment

The mixed-effects ordinal propensity score is derived from a mixed-effects ordinal logistic regression model developed by Hedeker and Gibbons (1994) in which a continuous latent variable ($\tau$) is related to an observed ordinal variable ($T$) through the threshold concept. If there are $K$ ordered categories of the observed outcome variable, the mixed-effects ordinal logistic regression
model assumes there are $K - 1$ strictly increasing thresholds, $\gamma_k$, ($k = 1, \ldots, K - 1$) on the scale of $\tau$ such that a response falls into an observed ordinal category $k$ if the latent variable $\tau$ is between the threshold values $\gamma_{k-1}$ and $\gamma_k$. (For example, $\tau$ may be a dose of a drug in milligrams and $T$ may be the dose categorized as high, medium, or low. In this case there are three ordered categories and therefore two thresholds, $\gamma_1$ and $\gamma_2$. The dose is categorized as medium on the ordinal scale if $\tau$ exceeds $\gamma_1$ but does not exceed $\gamma_2$.)

The model evaluates the effect of covariates and random effects on the $K - 1$ cumulative logits

$$\text{logit}[P(T_{ij} > k)] = \log \left[ \frac{P(T_{ij} > k)}{1 - P(T_{ij} > k)} \right] = \gamma_k + \nu_i + \beta' x_{ij},$$

for subject $i$ ($i = 1, \ldots, N$), at time $j$ ($j = 1, \ldots, n_i$), for dose $k$ ($k = 1, \ldots, K - 1$). The $x_{ij}$ are $p \times 1$ vectors of observed covariates, both time-varying and time-invariant. $\beta$ is a $p \times 1$ vector of fixed effects regression parameters. It is assumed that the subject-varying random effect $\nu_i$ is normally distributed with mean zero and variance, $\sigma^2_{\nu}$. As covariate effects are assumed to be the same across the cumulative logits, the model makes the proportional odds assumption (McCullagh, 1980).

The $p + K - 1$ regression parameters ($\gamma_1, \ldots, \gamma_{K-1}$ and $\beta$) are estimated by maximizing the marginal likelihood equations using Gauss-Hermite quadrature (Hedeker and Gibbons, 1994). The subject varying random effects are estimated using the posterior empirical Bayes estimate of the random effects.

Based upon this model, the \textit{mixed-effects ordinal propensity score} is defined as

$$e(x_{ij}, \nu_i) = P(T_{ij} > k | \nu_i, \tau) = \frac{1}{1 + \exp[-(\nu_i + \beta' x_{ij})]},$$
and is estimated by

\[ \hat{e}(x_{ij}, \nu_i) = P(T_{ij} > k|\hat{\nu}_i, x) = \frac{1}{1 + \exp[-(\hat{\nu}_i + \beta'x_{ij})]}. \]

where \( \hat{\beta} \) is a vector of estimated fixed effect regression parameters and \( \hat{\nu}_i \) is the empirical Bayes estimate of the random effect for subject \( i \). Inclusion of thresholds into the propensity scores is not necessary as the thresholds do not vary by subject or by time, and thus do not effect the rank ordering of propensity scores. Indeed, the analyst need only compute the estimated linear predictor

\[ \hat{\eta}_{ij} = \hat{\nu}_i + \beta'x_{ij} \]  

for each subject as the mixed-effects ordinal propensity score is a monotonic function of \( \hat{\eta}_{ij} \).

Once the mixed-effects ordinal propensity scores are estimated for each subject on each treatment occasion, observations are classified into propensity quintiles, \( q^{(1)}, \ldots, q^{(5)} \). A contingency table of treatment level by propensity quintile is constructed to determine whether each treatment level is well-represented in each propensity quintile. In addition, other statistical (Leon and Hedeker, 2011) and graphical techniques (Helmreich and Pruzek, 2009) can be used to assess the degree of covariate balance obtained within strata. If treatment levels are well-represented in all strata, and adequate covariate balance has been achieved, one can proceed with the stratified treatment effectiveness analysis.

### 2.2.2 Stratified Treatment Effectiveness Analysis

The next step in this analytic strategy is to conduct separate effectiveness analyses on observations in each propensity quintile. The type of analysis undertaken here depends on the nature of the response variable. In Leon et al. (2001), Leon and Hedeker (2005), and Leon et al. (2007), a mixed-effect grouped-time (survival) model for treatment effectiveness was
fitted for an outcome variable measuring time until recurrence of disease. In Leon and Hedeker (2007a) a mixed effects linear model was used to evaluate the effectiveness of treatment on a continuous outcome variable. In Leon and Hedeker (2011) a mixed-effects logistic regression model was used in the effectiveness analysis with a repeated binary outcome. Other models may be appropriate depending on the nature of the outcome variable, so long as they take into account the longitudinal nature of the data.

In the case of a longitudinal continuous outcome or response variable, say $y_{ij}$, with an ordinal treatment, a mixed-effects linear regression model (Laird and Ware, 1982) such as the following could be appropriate:

$$y_{ij} = \alpha_0 + \alpha_2 T_{ij2} + \ldots + \alpha_K T_{ijk} + \theta_i + \epsilon_{ij}$$

where $\alpha_0$ is the intercept term, $\alpha_k$ is the coefficient of the dummy-coded treatment dose $T_{ijk}$, $k = 2, \ldots, K$, $\theta_i$ is a subject-specific random intercept (distinct from $\nu_i$ in the random effects propensity model), and $\epsilon_{ij}$ is the error term.

Once the quintile-specific parameters of interest have been estimated, a test for interaction between treatment and propensity for treatment intensity is undertaken to examine whether the treatment effect varies across strata. In this stage of the analysis two models are fit, using the full, unstratified data set. A full model is fit that includes all the covariates included in the quantile-specific analysis, but the model also includes a propensity quintile classification
variable, treatment by propensity quintile interaction terms, and a subject-specific random effect. In the case of a continuous outcome variable, the full model would be

\[ y_{ij} = \alpha_0 + \alpha_2^* T_{ij2} + \ldots + \alpha_K^* T_{ijK} + \zeta_2 Q_2 + \ldots + \zeta_5 Q_5 + \delta_1 T_{ij2} Q_2 + \ldots + \delta_{(K-1)\times 4} T_{ijK} Q_5 + \theta_i^* + \varepsilon_{ij} \]

where \( \alpha_l, (l = 0, 2, \ldots, K) \), are the intercept and coefficients of the dummy treatment variables, \( \zeta_2, \ldots, \zeta_5 \), are the coefficients of their respective propensity quintile classification (dummy) variables, \( Q_2, \ldots, Q_5, \delta_1, \ldots, \delta_{(K-1)\times 4} \) are the coefficients of the propensity quintile by treatment level interactions, \( \theta_i^* \) is a subject-specific random effect and \( \varepsilon_{ij}^* \) is an error term. A reduced model is then fit that omits all treatment by propensity quintile interaction terms.

A likelihood ratio test comparing the full \((M_F)\) and reduced \((M_R)\) models is used to detect statistically significant treatment by propensity quintile interaction. Under the null hypothesis that \( \delta_1 = \ldots = \delta_{(K-1)\times 4} = 0 \), the test statistic

\[ \lambda = 2(\ell(M_F) - \ell(M_R)) \]

follows approximately a chi-square distribution with degrees of freedom equal to \((K - 1) \times 5\), where \( \ell(M_F) \) and \( \ell(M_R) \) denote the log likelihood evaluated at the ML estimators of the full and reduced models, respectively. If no significant interaction is detected, and one can safely assume homogeneity of treatment effectiveness across strata, then the quintile-specific parameter estimates are pooled using weighting procedure to compute an overall estimate of treatment effectiveness. In this procedure each quintile-specific parameter estimate is weighted by the inverse of its squared estimated standard error and the weighted estimates are summed.
The weighted pooled estimators of the treatment effectiveness parameters, $\alpha_\ell$, $\ell = 0, 2, \ldots, K$, are

$$\hat{\alpha}_\ell^p = \frac{\sum_{m=1}^{5} \frac{\hat{\alpha}_{\ell m}}{\text{se}(\hat{\alpha}_{\ell m})^2}}{\sum_{m=1}^{5} \text{se}^{-2}(\hat{\alpha}_{\ell m})}$$

where $\hat{\alpha}_{\ell m}$ is the $m$-th quintile-specific parameter estimate of $\alpha_\ell$ and $\text{se}(\hat{\alpha}_{\ell m})$ is the estimated standard error of the parameter estimate for the $m$-th quintile. Pooled standard errors are calculated in a similar fashion. Under the null hypothesis of no treatment effect, the statistic

$$T = \frac{\hat{\alpha}_\ell^p}{\sum_{m=1}^{5} \text{se}^{-2}(\hat{\alpha}_{\ell m})}$$

has an approximate standard normal distribution, which allows for the calculation of confidence intervals and p-values.

Leon and colleagues have published several simulation studies evaluating the performance of the mixed-effects quintile stratification propensity adjustment strategy. In a 2005 paper, the performance of the mixed-effects time-varying propensity adjustment with repeated survival intervals as outcomes was studied (Leon and Hedeker, 2005). In the simulation study, the adjustment strategy performed well, with acceptable type I error rates, decent statistical power in sample sizes of adequate size, and small bias in the estimates of treatment effectiveness. These authors conducted another simulation study to evaluate the bias reduction achieved using the modeling strategy with continuous outcomes (Leon and Hedeker, 2007a). They also examined the performance of four forms of quantile stratification: median-split, terciles, quartiles, and quintiles. The simulations showed that bias reduction increased monotonically with the number of quantiles used for stratification. The quartile and quintile stratification strategies using time-varying propensity scores typically removed 80 to 90 percent of the bias resulting from an unadjusted models. The impact of misspecification of the propensity model on the performance
of the quintile stratified propensity adjustment with continuous outcomes has been examined in another published Monte Carlo simulation study (Leon and Hedeker, 2007b). Simulated data were generated using a model for the propensity for ordinal treatment doses. The propensity was then estimated using a misspecified model that omitted certain confounding variables. Performance of the models was evaluated. Omission of time-varying confounders strongly associated with treatment and outcome had a deleterious impact on bias, coverage, and type I error.

2.3 Consequences of Misspecification of the Propensity Score

Most mixed-effects regression models, whether linear, generalized linear, or non-linear models, assume that the random effects are multivariate normally distributed (or univariate normal, in the case of random-intercept models). Most applications of these models also assume normality, as the estimation procedures encoded in most software packages are based upon that assumption. However, random effects are unobserved, latent variables, so the normality assumption is difficult to assess. In the case of generalized linear mixed models, the empirical Bayes estimators do not follow a normal distribution, even when the true random-effects distribution is normal (Litière et al., 2007).

Verbeke and Lesaffre (1997) proved the consistency and asymptotic normality of maximum likelihood estimators in linear mixed models that assume normality of the random effects, even when the models are misspecified with respect to the random-effects distribution. However, this proof applies to linear mixed models only. Litière and colleagues (2007) show that the maximum likelihood estimators of fixed effects parameters are consistent, even when the random-effects distribution is misspecified, but only if the parameters are equal to zero and are not included in the random-effects structure of the model. For example, in a logistic mixed-effects model with a random intercept only, the type I error rate for testing the null hypothesis of no treatment
effect will be preserved, as the parameter representing the treatment effect is not included in the random-effects structure of the model.

Beyond these theory-based results, most investigations of the robustness of generalized linear mixed models have been conducted via simulation studies. There is a wide range of opinions, and lively debate, regarding the effect of misspecification on estimation in generalized linear mixed-effects models based upon these studies. Neuhaus and colleagues (1992) investigated the effect of misspecification of the random-effects distribution for logistic random-effects models for binary data. In their simulation study these authors found that there is minimal bias in the estimation of the fixed effect parameters, even when the true random-effects distribution is highly skewed. However, under these circumstances, there is substantial bias in the estimation of the mean of the random-intercept distribution. Nevertheless, they found that the estimates of the standard errors of the fixed effects parameters estimates are relatively unbiased and concluded that inferences regarding the fixed effects are robust to misspecification of the random-effects distribution.

Agresti and collaborators (2004) investigated the effects of misspecification of the random-effects distribution on the efficiency of estimation in a generalized linear mixed-effects model for binary outcome data that assumes normality of random effects. In this simulation study, a clustered binary response variable was fit using a logit random-effects model with various distributions, including normal, uniform, exponential, discrete with two mass points, and degenerate at a single point (i.e., not random). They found a loss in efficiency and increase in bias in the estimation of the fixed effect parameters when the true distribution is a distinct mixture of distributions. In addition, they found that when the true random-effects distribution is a mixture that is quite non-normal, the prediction of random effects is adversely affected, leading to an increase in bias and loss of efficiency when compared to a nonparametric estimation approach.
that does not assume normality of the random effects. Agresti et al. (2004) recommended a strategy of using both a parametric and a nonparametric approach to examine the sensitivity of estimation results to distributional assumptions.

Litière and colleagues (2007) also conducted a simulation study using a binary response generated using a logistic random-intercept model. They found that misspecification severely affects the power of the analysis, depending on the shape and variance of the true random-effects distribution. Type I error rate, on the other hand, is quite robust to misspecification, even in extremely skewed distributions with large variances. Consistent with the theoretical results summarized above, Litière et al. (2007) found that the performance of the Wald test statistic to test the null hypothesis of no treatment effect is excellent even in small sample sizes ($n = 25$). However, the performance of the Wald test associated with the intercept parameter $\beta_0$ is very poor when the data-generating distribution is non-normal, even when the sample size is large ($n = 100$) and the random intercept variability is small relative to the treatment effect size.

Heagerty and Kurland (2001) evaluated the impact of misspecification of random effects on the maximum likelihood estimates based upon logistic random intercept regression models for clustered binary response data. These authors measured bias induced by misspecification by calculating the asymptotic relative bias, defined as: $100 \times (\theta^* - \theta_0)/\theta_0$, where $\theta^*$ is the value of the fixed effects parameter vector $\theta$ that minimizes the Kullback-Leibler information criteria and $\theta_0$ is the true fixed effects parameter vector.

Heagerty and Kurland (2001) considered four situations of random effects misspecification in their simulation study: (1) non-Gaussian random intercepts with a gamma distribution, (2) dependence of the random intercept variance on a cluster-level covariate, (3) omission of a Gaussian random slope, and (4) autocorrelated random effects. In agreement with Neuhaus et
al. (1992), they found that estimation of fixed effects parameters is robust to violations of the assumption of normality in the form of skewed distributions. Only highly skewed distributions with large between-cluster heterogeneity in the random effects lead to unacceptable levels of bias. Within-cluster parameter estimates and estimates of within-cluster and between-cluster interactions are seriously biased when the random intercept variance depends on a cluster-level covariate. Moderate bias in the estimated regression coefficients results when the random effects structure is misspecified by omitting a random slope for a within-cluster covariate. Finally, they found downward bias in the absolute magnitude of the estimated regression coefficients. Heagerty and Kurland (2001) recommended careful attention to the random effects model assumptions.

Yucel and Demirtas (2010) investigated the effect of misspecification of the random effects distribution on the performance of multiple imputation for missing data in the context of a multivariate linear mixed-effects model via a simulation study. They found that under most conditions the multiple imputation procedure is fairly robust to misspecification in the estimation of the fixed effects parameters of the model. However, the estimation of the variance of the random effects is far more inconsistent, especially under the circumstances where both the predictor and response variables are subject to fairly high rates of missingness. The authors concluded that inferences regarding the random effect variance based upon the multiple imputation procedure when missingness is present in both response and predictor variables are "highly questionable". (However, it should be noted that in virtually every simulation scenario the multiple imputation procedure performs far better than the method of restricting the analysis to completely observed cases only.)

Finally, Verbeke and Lesaffre (1996) demonstrated that empirical Bayes estimates of the random effects can suffer severe distortion when the assumption of normality is incorrect.
In summary, although some authors claim that misspecification of the random effects distribution has a limited impact on the estimation of fixed effects, most authors agree that estimates of fixed effects in generalized linear mixed models are adversely affected when the true random effects distribution is highly skewed or bimodal. The impact of misspecification on estimation of fixed effects in GLMM is enhanced when the random effects are highly variable. Finally, almost all agree that misspecification of the random effects distribution adversely affects inference regarding the random effects. It should be noted, however, that all of the studies evaluating the impact of random effects misspecification on estimation in generalized linear mixed models deal with the situation of binary response variables only. None, so far as this author is aware, have evaluated directly the sensitivity of ordinal mixed effects models to distributional assumptions regarding the random effects.

The mixed-effects ordinal propensity score that is used in the dynamic propensity adjustment approach for longitudinal observational data is derived from a mixed-effects ordinal logistic regression model that assumes normality of the random effects distribution. The mixed effects propensity score is obtained by evaluating the estimated regression equation, which is a function of the maximum likelihood estimate of the fixed effects, $\hat{\beta}$, and the empirical Bayes estimate of the random effect for the $i$th subject, $\hat{\nu}_i$. If the estimation of either or both of these is sensitive to misspecification of the unobservable random effects distribution, then it is reasonable to assume that the quality of the dynamic propensity adjustment may be negatively impacted as well. Hence, the interest in more flexible estimation approaches for mixed effects ordinal regression which relax the normality assumption.

2.4 Robust Alternatives to the Normal Model

Given the wide range of opinion regarding the impact of misspecification of the random effects distribution, and the unavailability of methods to check the soundness of the normality
assumption, it is not surprising that alternative models have been developed that are robust to violations of this distributional assumption.

Butler and Louis (1992) applied the technique of non-parametric maximum likelihood estimation (NPML) of a general distribution function, developed by Laird (1978), to the estimation of random effects model parameters for both continuous and discrete data. In this approach the random effects distribution is approximated by a discrete distribution with mass at a finite number of points. The E-M algorithm (Dempster et al., 1977) is employed to find the ML estimates. Butler and Louis (1992) found in a simulation study that the NPML estimates are robust to departures from a Gaussian random effects distribution while remaining efficient in the case of Gaussian random effects. The technique, however, is limited to a model with a single random effect as models with multivariate random effects create identifiability problems without strong assumptions.

Drawing on the work of Laird and Louis (1991), Shen and Louis (1999) proposed an iterative method of obtaining a smooth estimate of the random effects distribution in a linear mixed effects model, which they called the smoothing by roughening (SBR) approach. They begin with a initial discrete approximation of a continuous distribution on a equally spaced grid of points. The recursive method converges to an approximate NPML with the quality of the approximation depending on the fineness of the grid chosen. Let \( b = (b_1, \ldots, b_M) \) be a grid of \( M \) evenly spaced mass points and \( f^{(0)}(b) = (f^{(0)}(b_1), \ldots, f^{(0)}(b_M)) \) be the probability at each respective mass point under the initial discrete density. The recursive process updates the distribution of \( b \) as follows:

\[
f^{(r+1)}(b_m) = f^{(r)}(b_m) \frac{1}{N} \sum_{i=1}^{N} \frac{f(Y_i|b_m)}{\sum_{k=1}^{M} f(Y_i|b_k) f^{(r)}(b_k)},
\]
where \( f(Y_i|b_m) \) is the conditional distribution of \( Y_i \) given \( b_m \), \( m = 1, \ldots, M \). For the linear mixed effects model, \( f(Y_i|b_m) = \phi(X_i\beta + Z_i b_m, \sigma^2 I_{n_i}) \), where \( \phi \) is the pdf of the normal distribution. The discretized random effects distribution is reiteratively updated until convergence.

Verbeke and Lesaffre (1996) developed a linear mixed-effects model with heterogeneity in the random-effects population to cope with non-normality of random effects. The model, which is an extension of the classical random effects model, assumes that the random effects are sampled from a finite mixture of \( G \) normal distributions with means \( \mu_g \) and covariance matrix \( D_g \), \( g = 1, \ldots, G \). The random effects \( b_i \) are distributed as the weighted sum of \( g \) normal distributions, where the weights are the probability of being sampled from a given component of the mixture, \( p_g \), where \( \sum_{g=1}^{G} p_g = 1 \). The outcome variable, \( y_i \) is then distributed as a weighted sum of \( G \) normal distributions with mean \( X_i\beta + z_i \mu_g \) and covariance matrix \( V_i \), where \( V_i = z_i D z_i^T + \sigma^2 I_{n_i} \). The constraint is added that \( E(b_i) = \sum_{g}^{G} p_g \mu_g = 0 \) to ensure that \( E(y_i) = X_i\beta \).

The EM algorithm can be used to obtain maximum likelihood estimators for all parameters, and empirical Bayes estimates can be used to estimate the posterior probability of the \( i \)th individual belonging to the \( g \)-th component of the mixture. In practice, it is usually assumed that all \( D_g = D \), i.e., that the covariance matrices of all components of the mixture are the same.

Fieuws et al. (2004) have developed a SAS macro to estimate parameters of the heterogeneity model using the SAS procedure NLMIXED (SAS Institute, 2017). Litière and collaborators (2006) have extended the heterogeneity model to generalized linear mixed models and developed another SAS macro based on the NLMIXED procedure to estimate these models. As we propose to employ the heterogeneity model for GLMMs in our dynamic propensity for treatment model outlined below, we will defer an extended consideration of this model and estimation of model parameters until the discussion of our proposed model.
Litière and colleagues (2006) also conducted a simulation study of the heterogeneity model for generalized linear mixed models to assess the performance of this model relative to the standard homogeneity model. They generated binary longitudinal data using a random-intercept logistic model with a treatment effect and a fixed effect for time. The random effects in this study were generated using a variety of distributions, including normal, log-normal, uniform, power function and asymmetric mixture of two normals. The sample size was also varied. The generated data were fit to a standard GLMM and to a heterogeneity model assuming a mixture of two normal distributions. The authors found that the heterogeneity model for GLMMs generally performed "somewhat better" in terms of bias of fixed effects parameter estimates than the homogeneity model when the random effects came from highly skewed or heavy-tailed distributions, especially when sample sizes were small (n = 50). However, the authors also found that the bias reduction came at a cost: the heterogeneity model frequently did not converge and the estimates were sensitive to the choice of starting values for the parameter estimates. (See also Komárek (2001) and Fieuws, Spiessen and Draney (2004).)

Ghidey et al. (2004) proposed a linear mixed model with a smooth random effects density using penalized spline smoothing, based upon the work of Eilers and Marx (1996). This approach combines a flexible modeling of the random effects distribution using a weighted mixture of Gaussian basis functions with a penalty for over-smoothing. Following Eilers and Marx (1996), they use a penalty based upon second-order finite differences between coefficients of adjacent basis functions, thus avoiding the need to calculate first and second derivatives during the estimation procedure. A Newton-Raphson optimization algorithm is used to maximize the penalized marginal density. Ghidey et al. (2010) showed that the penalized Gaussian mixed model (PGM) produces consistent parameter estimates.
Ghidey et al. (2010) compared the performance of the aforementioned PGM model with other methods used to accommodate non-normal random effects distributions, including Shen and Louis’ (1999) SBR approach and the heterogeneity model of Verbeke and Lesaffre (1996). They concluded from their simulation studies that the PGM method performed better when the underlying random effects distribution was highly skewed or heavy tailed, while the heterogeneity model performed better when the true random effects distribution was a mixture. The SBR approach performed adequately but did not outperform the other approaches, except in limited circumstances.

Komárek and Lesaffre (2008) have extended the PGM approach of Ghidey (2004) to generalized linear mixed models. Although they have developed the model to handle multiple random effects, for simplicity of presentation, we will only present the random intercept model.

In the PGM GLMM, we assume that the random intercept, \( b \) is distributed with an unknown density, \( g(b) \), and

\[
g(b) = \frac{1}{\tau} g^*(b^*)
\]

where \( \tau \) is an unknown scale parameter and \( b^* \) is a standardized random effect. A fine grid of \( L = 2K + 1 \) of equidistant knots, \( \mu = (\mu_{-K}, \ldots, \mu_K)' \) is chosen and centered at zero, with \( \mu_0 = 0 \). If the distance between two consecutive knots is \( \delta \), then \( \mu_j = j\delta, j = -K, \ldots, K \). A basis standard deviation \( \sigma \) is chosen and the model for \( g^* \) is:

\[
g^*(b^*) = \frac{1}{\sigma} \sum_{j=-K}^{K} w_j \phi \left( \frac{b^* - \mu_j}{\sigma} \right)
\]

where \( \phi \) is the standard normal density and \( \{w_j : j = -K, \ldots, K\} \) is a set of \( L \) mixture weights. \( g^*(\cdot) \) is a mixture of univariate normal densities with the sequence of knots serving as their means. The knots are placed in the area of high probability mass for a mean-zero, unit
variance density. Komárek and Lesaffre (2008) suggest $\mu_{-K} = -5$ and $\mu_K = 5$ and that $\delta$, the distance between knots, equal to 0.3, to approximate $g^*(\cdot)$ adequately. Finally, the standard deviation for the basis functions is set to $\sigma = \left(\frac{2}{3}\right)\delta$.

In order to avoid constrained optimization, Komárek and Lesaffre recommend the use of transformed weights:

$$a_j = \log \left( \frac{w_j}{w_0} \right), \quad w_j(a) = \frac{\exp(a_j)}{\sum_{l=-K}^{K} \exp(a_l)}$$

where $w_0$ is the weight at $\mu_0 = 0$.

For computational convenience, Komárek and Lesaffre use a Bayesian approach to estimation and inference, although they emphasize that a likelihood approach is possible, if computationally demanding. Non-informative priors on all model parameters are chosen. Sampling from the posterior distribution is accomplished with a hybrid Gibbs sampler. For computational details, we refer to Komárek and Lesaffre (2008). The authors have also developed an R contributed package called glmmAK that will produce MCMC simulations and posterior calculations for several types of GLMMs (Komárek, 2012).

Komárek and Lesaffre (2008) also conduct a small simulation study to test the performance of the PGM GLMM relative to the normal GLMM with binary outcome data. They find that in most circumstances, the PGM GLMM outperformed the normal GLMM when the random intercept of the underlying logit model was generated either from a log-normal (skewed) distribution or a normal mixture. Moreover, the decrease in bias in the estimation of model parameters was achieved at only a modest price in terms of an increase in the variability of the estimates. To our knowledge, a head-to-head comparison between the heterogeneity model for GLMMs and the PGM GLMM has yet to be conducted.
CHAPTER 3

PROPOSED MODEL

Let $Y_{ij}$ be a continuous response of interest for subject $i$ at measurement occasion $j$, ($i = 1, \ldots, N$), and ($j = 1, \ldots, n_i$), and let $Y_i$ be the $n_i$-dimensional vector of measurements on response variable $Y$ for the $i$th subject. Further, let $f(Y_i|\nu_i)$ be the conditional density of $Y_i$ given the random intercept for the $i$th subject, $\nu_i$. Assume that the primary explanatory variable of interest is an ordinal variable, $T_{ij}$, that can assume $K$ levels: $T_{ij} \in \{1, 2, \ldots, K\}$. Assume further that a set of explanatory and potentially confounding variables is measured on each measurement occasion. The vector of covariates for each measurement occasion $j$ for the $i$th subject is denoted by $x_{ij}$.

3.0.1 Mixed-effects Ordinal Regression Model for Propensity for Treatment

In what follows, we follow closely the discussion of the heterogeneity model for generalized linear mixed models as outlined in Molenberghs and Verbeke (2005, chap. 23).

3.0.2 Propensity for Treatment Model - The Heterogeneity Model

As mentioned in the earlier discussion of the dynamic propensity model approach, the mixed effects random intercept model for the propensity for treatment is:

$$\logit [P(T_{ij} > k|x_{ij}, \beta, \nu_i)] = \gamma_k + \nu_i + \beta'x_{ij}, \quad (3.1)$$

where $\gamma_k$ are $K - 1$ strictly increasing thresholds ($\gamma_1 < \gamma_2 \ldots < \gamma_{K-1}$), and $\gamma_1$ is set to zero for identifiability of the threshold parameters. As mentioned previously, since the fixed effects parameters do not vary by category $k$, this model assumes that the effects of those
parameters does not depend on $k$. This is commonly known as the proportional odds assumption (McCullagh, 1980).

Rather than assume that the random intercept in this model is normally distributed across the population of subjects, we will assume that the subjects are sampled from a mixture of $g$ component subpopulations, each with an associated component distribution. Each of these component distributions is normally distributed, and the $r$th component is distributed as normal with mean $\mu_r$ and variance $\sigma_r^2$. That is:

$$\nu_i \sim \sum_{r=1}^{g} p_r N(\mu_r, \sigma_r^2), \quad (3.2)$$

where $p_r$ is the prior probability that $\nu_i$ is sampled from component $r$ and $\sum_{r=1}^{g} p_r = 1$. Note that we will assume that the variance of the random intercept term does not vary across component subpopulations.

Define the indicator variable $Z_{ir}$, which is equal to 1 if $\nu_i$ is sampled from the $r$th component of the mixture, and equal to 0 otherwise, $r = 1, \ldots, g$. That is, each subject is associated with a $(g \times 1)$ vector $z_i$, such that the $r$th component of $z_i$ is equal to 1, and the other components are equal to zero, if the subject comes from the $r$th component population.

Note that $P(Z_{ir}) = 1 = E(Z_{ir})$, which we will denote $p_r$.

$$E(\nu_i) = E[E[\nu_i | z_{i1}, \ldots, z_{ig}]]$$

$$= E \left[ \sum_{r=1}^{g} \mu_r z_{ir} \right]$$

$$= \sum_{r=1}^{g} p_r \mu_r$$
Thus, in order to assure that $E[\nu_i] = 0$, we need to add the additional constraint that 

$$\sum_{r=1}^{g} p_r \mu_r = 0.$$ 

The variance of $\nu$ is derived as follows:

$$\sigma^2_{\nu} = \text{var} \left[ E(\nu_i|z_{i1}, \ldots, z_{ig}) \right] + E\left[ \text{var}(\nu_i|z_{i1}, \ldots, z_{ig}) \right]$$

$$= \text{var} \left( \sum_{r=1}^{g} \mu_r z_{ir} \right) + E \left( \sum_{r=1}^{g} \sigma^2_{\nu} z_{ir} \right)$$

$$= \sum_{r=1}^{g} p_r \mu_r^2 + \sum_{r=1}^{g} p_r \sigma^2_{\nu}$$

Let $f_r(\nu_i)$ be the $r$th mixture component density. The prior density of $\nu_i$ is

$$f(\nu_i) = \sum_{r=1}^{g} p_r f_r(\nu_i)$$

$$= \sum_{r=1}^{g} p_r (2\pi \sigma^2_{\nu})^{-\frac{1}{2}} \times \exp \left\{ -\frac{1}{2} \left( \frac{\nu_i - \mu_r}{\sigma_{\nu}} \right)^2 \right\}$$

The number of components in the mixture, $g$, is considered known.

### 3.0.3 Propensity for Treatment Model - Estimation and Inference

The maximum marginal likelihood approach is used for estimation and inference.

$$f_i(y_i) = \int f_i(y_i|\nu_i) f(\nu_i) d\nu_i$$

$$= \sum_{r=1}^{g} p_r \int f_i(y_i|\nu_i) f_r(\nu_i) d\nu_i$$

$$= \sum_{r=1}^{g} p_r f_{ir}(y_i),$$

where $f_{ir}(y_i)$ is the marginal density of a mixed model with a random intercept that is normally distributed with mean $\mu_r$ and variance $\sigma^2_{\nu}$. The marginal density of $Y_i$ is a $g$-component finite mixture.
Molenberghs and Verbeke (2005) suggest maximizing the marginal likelihood using the E-M algorithm (Dempster et al., 1977). The E-M algorithm is often employed to tackle missing data problems. In the heterogeneity model, the group component labels \( Z_{ir} \) and thus the mixing weights \( p_r \) are unknown and treated as missing. In general, given a parameter vector \( \theta \) and observed data \( x \), the E-M algorithm attempts to maximize the likelihood \( L(\theta|x) \) with respect to \( \theta \). The process begins with an initial estimate of the parameters, \( \theta^{(0)} \), from which one computes the expectation of the joint log likelihood of the complete data \( Q(\theta|\theta^{(0)}) \) (the E step):

\[
Q(\theta|\theta^{(0)}) = E \left\{ \log L(\theta|Y)|x,\theta^{(0)} \right\} = E \left\{ \log f(Y|\theta)|x,\theta^{(0)} \right\}
\]

In the first iteration of the M step, we maximize \( Q(\theta|\theta^{(0)}) \) with respect to \( \theta \). We then set \( \theta^{(1)} = \arg \max_\theta Q(\theta|\theta^{(0)}) \), and return to the E step and compute \( Q(\theta|\theta^{(1)}) \), and so on until convergence (Givens and Hoeting, 2012).

Some additional notation is necessary to develop the expectation of the conditional log likelihood for the heterogeneity model. The vector \( \pi = (p_1, \ldots, p_g) \) is the vector of component probabilities; the vector of the remaining parameters in the density \( f_i(y_i|\nu_i) \) is \( \xi = (\beta', \mu_r, \sigma_r^2)' \).

Finally, all parameters in the marginal heterogeneity model are included in \( \theta = (\pi', \xi')' \).

The marginal likelihood function is:

\[
L(\theta|y) = \prod_{i=1}^{N} \left\{ \sum_{r=1}^{g} p_r f_{ir}(y_i|\xi) \right\},
\]

(3.3)
where \( \mathbf{y}' = (y'_1, \ldots, y'_N) \) is the vector containing all observed responses. The joint log likelihood for the observed outcome measurements \( \mathbf{y} \) and the vector \( \mathbf{z} \) of unobserved component assignments, \( z_{ir} \), is

\[
\ell(\theta | \mathbf{y}, \mathbf{z}) = \sum_{i=1}^{N} \sum_{r=1}^{g} z_{ir} [\ln p_r + \ln f_{ir}(y_i | \xi)] ,
\]

where \( p_r = P(Z_{ir} = 1) \).

In the E-step, we find the conditional expectation of the joint log likelihood, which is:

\[
Q(\theta | \theta^{(t)}) = E[\ell(\theta | \mathbf{y}, \mathbf{z}) | \mathbf{y}, \theta^{(t)}] \\
= \sum_{i=1}^{N} \sum_{r=1}^{g} p_{ir}^{(\theta^{(t)})} [\ln p_r + \ln f_{ir}(y_i | \xi)] \\
= \left( \sum_{i=1}^{N} \sum_{r=1}^{g} p_{ir}^{(\theta^{(t)})} \ln p_r \right) + \left( \sum_{i=1}^{N} \sum_{r=1}^{g} p_{ir}^{(\theta^{(t)})} \ln f_{ir}(y_i | \xi) \right)
\]

where the expectation in Equation 3.5a is taken over the missing data \( \mathbf{z} \), and the expression \( p_{ir}^{(\theta^{(t)})} \) is used to make explicit the dependence of \( p_{ir} \) on the current parameter estimates \( \theta^{(t)} \) and

\[
p_{ir}^{(\theta^{(t)})} = E(z_{ir} | \mathbf{y}_i, \theta^{(t)}) \\
= P(z_{ir} = 1 | \mathbf{y}_i, \theta^{(t)}) \\
= \frac{p_r f_{ir}(y_i | \xi)}{\sum_{k=1}^{g} p_k f_{ik}(y_i | \xi)}
\]

In the maximization or M-step, we need to maximize Equation 3.5c with respect to \( \theta \). Molenberghs and Verbeke suggest that we separately maximize the two summands (in large
parentheses) in Equation 3.5c separately, and that we further decompose the first summand as follows:

$$
\sum_{i=1}^{N} \sum_{r=1}^{g} p_{ir}^{(\theta_{(i)})} \ln p_r = A + B,
$$

(3.7)

where A and B are as follows:

$$
A = \sum_{i=1}^{N} \sum_{r=1}^{g-1} p_{ir}^{(\theta_{(i)})} \ln p_r
$$

(3.8)

$$
B = \sum_{i=1}^{N} \left[ p_{ig}^{(\theta_{(i)})} \ln \left( 1 - \sum_{r=1}^{g-1} p_r \right) \right]
$$

(3.9)

We maximize Equation 3.7 with respect to $p_1, \ldots, p_{g-1}$. Molenbergh and Verbeke show that the updating estimates satisfy

$$
p_{r+1}^t = \frac{1}{N} \sum_{i=1}^{N} p_{ir}^{(\theta_{(i)})},
$$

where $p_{ir}^{(\theta_{(i)})}$ is given above in Equation 3.6c.

The second summand in Equation 3.5c cannot be maximized analytically. However, Molenberghs and Verbeke point out that the second summand can be interpreted as a weighted log-likelihood of a generalized linear mixed model. Maximization can be achieved using procedures for fitting generalized linear models, such as SAS procedure NLMIXED (SAS Institute, 2017) or R packages nlme (Pinheiro et al., 2017) or lme4 (Bates et al., 2017).

### 3.0.4 Propensity for Treatment Model - Empirical Bayes Estimation

In order to estimate the propensity for treatment for subject $i$ at measurement occasion $j$, we need to obtain the empirical Bayes estimate (EBE) of the subject-specific random intercept, $\nu_i$. The EBE of $\nu_i$ is given by

$$
\hat{\nu}_i = \sum_{r=1}^{g} p_{ir}^{(\theta_{(i)})} \hat{\nu}_{ir}^{(\xi)},
$$

(3.10)
which is the weighted sum of the component-specific posterior modes with \( \theta \) and \( \xi \) replaced with their maximum likelihood estimates.

### 3.0.5 Propensity for Treatment Model - Ranking the Propensity for Treatment

Recall that the dynamic propensity score adjustment procedure employs a stratification strategy to adjust for bias introduced by confounding. In order to obtain the ranks of the propensity for treatment score, as discussed in Section 2.2.1 on page 10, we need only compute

\[
\hat{\eta}_i = \hat{\nu}_i + \hat{\beta}'x_{ij}
\]

Once the mixed-effects ordinal propensity scores are estimated for each subject on each measurement/treatment occasion, observations are classified into propensity quintiles, \( q^{(1)}, \ldots, q^{(5)} \). If the covariate balance achieved is adequate, then the analysis proceeds to the stratified treatment effectiveness analysis.

### 3.0.6 Stratified Linear Mixed-Effects Model for Effectiveness

The strategy for estimating treatment effectiveness, adjusting for the propensity for treatment, is the same as developed by Leon and Hedeker (2007b) and as outlined in Section 2.2.2 above:

1. Fit a stratified linear mixed effects model with each observation stratified according to the estimated propensity for treatment intensity quintile for subject \( i \) at treatment occasion \( j \), as estimated in the propensity for treatment intensity model.

2. Test for interaction between treatment and propensity for treatment intensity quintile to determine whether treatment effect varies across treatment intensity quintiles.
(a) Fit a full model ($M_F$) that includes all the covariates included in the quantile-specific analysis and that also includes a propensity quantile classification (dummy) variable, treatment by propensity quintile interaction terms, and a subject-specific random intercept term.

(b) Fit a reduced model ($M_R$) that omits all treatment by propensity intensity quintile interaction terms in the full model.

(c) Perform a likelihood ratio test comparing the full and reduced models to detect statistically significant treatment by propensity intensity quintile interaction.

   i. If the null hypothesis of no treatment by treatment intensity quintile is rejected, then the results of the stratum-specific analyses cannot be pooled, and further analysis is reasonable only within treatment intensity strata only.

   ii. If the null hypothesis is rejected, then proceed to the next step.

3. Given homogeneity of treatment effectiveness across treatment intensity strata, pool the quintile-specific parameter estimates using the Mantel-Haenzel procedure to compute an overall estimate of treatment effectiveness.

CHAPTER 4

SIMULATION STUDY

4.1 Overview

In order to examine the impact of non-normality of the propensity for treatment intensity on the performance of Leon and Hedeker’s dynamic propensity model adjustment for longitudinal data and to test the performance of the proposed modification to that procedure, Monte Carlo simulation studies were conducted. The primary focus was a comparative assessment of the performance of these models when the distribution for random intercept $\nu_i$ departs from normality. As outlined in more detail below, we employed linear random effects regression models with random effects drawn from varying distributions ($M_{Treat}$) to generate a latent, continuous treatment intensity for each subject at each treatment occasion. Using arbitrary threshold values, the latent treatment intensity was mapped onto four ordinal treatment dose levels. A linear random effects regression model ($M_{Resp}$) was used to generate continuous outcomes. The latter model used the same fixed effects as were used in $M_{Treat}$ in order to incorporate the phenomenon of confounding. The performance of the original dynamic propensity adjustment model ($DPAM-Norm$) and the model incorporating possible heterogeneity in the random effects ($DPAM-Het$) were compared. Performance was evaluated with respect to bias and precision of estimation of the parameter for treatment effectiveness, $\beta_T$. 

32
4.2 Simulation Specifications

The simulated data was generated in the following manner. A vector of predictor variables was generated for each subject at each treatment occasion:

\[ \mathbf{x}' = (x_1, x_2, x_3, x_4) \]

- \( x_1 \) is a time-invariant, continuous variable.
- \( x_2 \) is a time-invariant, binary variable.
- \( x_3 \) is a time-varying, continuous variable.
- \( x_4 \) is a time-varying, binary variable.

The random variables are generated using the \texttt{R} (R Core Team, 2017) package \texttt{BinNor}, which implements a methodology developed by Demirtas and Doganay (2012) for the simultaneous generation of binary variables and normal variables with specified marginal means and correlations (Amatya and Demirtas, 2016; Demirtas et al., 2014). A correlation is imposed on the set of predictors and \( \rho_{xx'} \) set to 0.20. The correlation between measurement occasions for each time-varying predictor \((x_3, x_4)\) is set to 0.40.

4.2.1 Distribution of Random Effects

Three different distributions for \( \nu_i \) were used to generate random intercepts:

1. A normal distribution, \( \nu_i \sim N(0, \sigma^2_\nu) \);
2. A zero-mean shifted log-normal distribution; and
3. A mixture of two normal distributions that is asymmetric and bimodal.

Distribution (1) was used to check the performance of the \textit{DPAM-Het} model under circumstances in which the \textit{DPAM-Norm} should outperform it. Distribution (2) is a highly skewed
distribution whose support does not cover the whole real line. Distribution (3), as mentioned, is skewed and bimodal.

4.2.2 Model for Propensity for Treatment ($M_{T \text{real}}$)

The odds ratios for each of the four predictors of dose in the propensity model were varied (1.25, 1.50, 1.75, 2.0). A latent, continuous dose was generated using a linear mixed effects model:

$$
\lambda_{ij} = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \nu_i + \varepsilon_{ij},
$$

(4.1)

where $\varepsilon_{ij}$ is distributed as a standard logistic, with variance $\frac{\pi^2}{3}$, i.i.d., and $\sigma^2_\varepsilon$ is set such that the intraclass correlation between repeated doses, $\rho_\lambda$, is equal to 0.40. That is

$$
\rho_\lambda = \frac{\sigma^2_\nu}{\sigma^2_\nu + \sigma^2_\varepsilon} = 0.40.
$$

The association of each covariate ($x_1 - x_4$) on both the continuous effectiveness outcome and dose was specified to be approximately equivalent. The observed time-varying ordinal dose, $T_{ij}$, was then generated using the threshold concept:

$$
T_{ij} = \begin{cases} 
1 & \text{if } -\infty < \lambda_{ij} \leq \delta_1 \\
2 & \text{if } \delta_1 < \lambda_{ij} \leq \delta_2 \\
3 & \text{if } \delta_2 < \lambda_{ij} \leq \delta_3 \\
4 & \text{if } \delta_3 < \lambda_{ij} < \infty
\end{cases}
$$

where $\delta_1 < \delta_2 < \delta_3 < \delta_4$. 

4.2.3 Model for Treatment Effectiveness ($M_{Resp}$)

A continuous outcome is generated using a linear mixed effects model:

$$Y_{ij} = \beta_0 + \beta_r \tau_{ij} + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \zeta_i + \epsilon_{ij},$$  \hspace{1cm} (4.2)

where the error term $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$, i.i.d., the random intercept $\zeta_i \sim N(0, \sigma^2_\zeta)$, i.i.d., $\beta_0$ is set to zero (indicating that the lowest ordinal dose level has no effect on response), and

$$\beta_r = (\beta_{r2}, \beta_{r3}, \beta_{r4}) = (0.11, 0.32, 0.66),$$  \hspace{1cm} (4.3)

(in standard deviation units) and $\tau_{ij}$ is a $3 \times 1$ dummy vector for the ordinal doses, defined as follows:

$$\tau_{ij} = \begin{cases} (0, 0, 0)' & \text{if } T_{ij} = 1 \\ (1, 0, 0)' & \text{if } T_{ij} = 2 \\ (0, 1, 0)' & \text{if } T_{ij} = 3 \\ (0, 0, 1)' & \text{if } T_{ij} = 4. \end{cases}$$

The variance of the error terms and the random intercept were chosen such that the ICC for repeated outcomes is 0.40.

4.2.4 Simulation Size

A sample size of 250 subjects was generated for each simulated data set and each subject had 8 repeated observations. One thousand data sets were generated for each combination of simulation specifications. Each data set was analyzed separately using the DPAM-Norm model and the DPAM-Het model.
4.3 Evaluation of Model Performance

The parameters of interest are the model-based estimates of the elements of $\beta_T$, the effects of ordinal treatment doses on outcome. Performance of the models were evaluated using the following criteria in the estimation of $\beta_T$: coverage rate, statistical power, standardized bias, and root mean square error (RMSE) (Demirtas, 2004; Demirtas, 2005; Demirtas et al., 2007; Demirtas and Hedeker, 2008).

4.4 Simulation Study Results

The results of the simulations conducted are summarized in Table I through Table III.

4.4.1 Normal Random Effects

Table I contains the results for the simulated data sets in which the random effects are normally distributed, consistent with the assumption of the generalized linear mixed effects model. The $DPAM-Norm$ and $DPAM-Het$ models exhibit very similar performance on all measures of bias and precision. This pattern is repeated across all levels of covariate effect size and all treatment levels. The result is not surprising since the two methods generated equivalent predictive propensity models in most cases. The normal model slightly outperforms the heterogeneity model in terms of standardized bias at all treatment levels and all covariate effect sizes. This result is also not surprising given that the heterogeneity model introduces an additional source of error in that it includes an additional model selection step in the determination of the magnitude of $g$, the number of components in the mixture distribution.

In general, the performance of both estimation methods deteriorates as the magnitude of the covariate/confounding variable effect size increases. Both methods underestimate the true effect size, when the log odds ratio of the covariate/confounding variable is greater than 1.75. Coverage is adequate at all factor levels. However, power is at acceptable levels at only the highest of treatment levels.
Finally, the root mean square error for both methods are virtually identical, indicating a similar overall level of accuracy and precision for both methods.

### 4.4.2 Non-normal Random Effects

The results of the simulations in which the true propensity model was generated from a normal mixture are shown in Table II and those simulations in which the random effects were generated from a log normal distribution are shown in Table III. Plots of the random effect distributions are shown in Figure 1 and Figure 2.

Given the results summarized in the literature review, it was expected that the performance of the two methods would diverge as the random effect distributions departed from the assumption of normality. This expectation was not fulfilled. The *DPAM-Norm* model proved to be fairly robust to departures from the assumption of random effects normality, even when the departures were relatively extreme. Due, in part, to the robustness of the *DPAM-Norm* model, the *DPAM-Het* model did not perform better than the *DPAM-Norm* model under most
TABLE II

NORMAL MIXTURE RANDOM EFFECTS

<table>
<thead>
<tr>
<th>Odds Ratio</th>
<th>Covariate</th>
<th>Method</th>
<th>Coverage</th>
<th>Power</th>
<th>SB</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Effect Size</td>
<td>Effect Size</td>
<td>Effect Size</td>
<td>Effect Size</td>
</tr>
<tr>
<td>1.25</td>
<td>Norm</td>
<td>.11</td>
<td>.32</td>
<td>.66</td>
<td>.11</td>
<td>.32</td>
</tr>
<tr>
<td></td>
<td>Hetero</td>
<td>.93</td>
<td>.95</td>
<td>.95</td>
<td>.19</td>
<td>.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.94</td>
<td>.95</td>
<td>.96</td>
<td>.19</td>
<td>.80</td>
</tr>
<tr>
<td>1.50</td>
<td>Norm</td>
<td>.95</td>
<td>.96</td>
<td>.95</td>
<td>.15</td>
<td>.74</td>
</tr>
<tr>
<td></td>
<td>Hetero</td>
<td>.95</td>
<td>.96</td>
<td>.95</td>
<td>.15</td>
<td>.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.94</td>
<td>.95</td>
<td>.95</td>
<td>.18</td>
<td>.67</td>
</tr>
<tr>
<td>1.75</td>
<td>Norm</td>
<td>.94</td>
<td>.95</td>
<td>.95</td>
<td>.18</td>
<td>.67</td>
</tr>
<tr>
<td></td>
<td>Hetero</td>
<td>.94</td>
<td>.95</td>
<td>.94</td>
<td>.17</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.94</td>
<td>.96</td>
<td>.94</td>
<td>.17</td>
<td>.68</td>
</tr>
<tr>
<td>2.00</td>
<td>Norm</td>
<td>.95</td>
<td>.93</td>
<td>.96</td>
<td>.17</td>
<td>.64</td>
</tr>
<tr>
<td></td>
<td>Hetero</td>
<td>.95</td>
<td>.94</td>
<td>.95</td>
<td>.17</td>
<td>.64</td>
</tr>
</tbody>
</table>

simulation scenarios explored in this study, and, in fact, sometimes performed worse. As the performance of both models was similar whether the random effects in the true propensity model were from a mixture distribution or log normal distribution, the results of these simulations will be discussed together.

With both a mixture of normals and with log normally distributed random effects in the data generating model for the true propensity score, there was very little discrepancy between the true coverage probabilities and the nominal coverage probabilities at all levels of simulation factors. Interestingly, the statistical power of the estimates of treatment effect at different levels actually improved when the propensity model random effects were non-normal. The standardized bias of the estimates did not show any consistent pattern; however, overall, the performance of both models was similar and generally acceptable. The estimate of the treatment level with effect size = 0.32 actually improved substantially, from the perspective of bias, at the highest covariate/confounder effect size level ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = \log(2.00)$). With
<table>
<thead>
<tr>
<th>Odds Ratio</th>
<th>Method</th>
<th>Covariate</th>
<th>Coverage Effect Size</th>
<th>Power Effect Size</th>
<th>SB Effect Size</th>
<th>RMSE Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>Norm</td>
<td>0.94</td>
<td>0.96</td>
<td>0.95</td>
<td>0.18</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Hetero</td>
<td>0.95</td>
<td>0.95</td>
<td>0.97</td>
<td>0.18</td>
<td>0.80</td>
</tr>
<tr>
<td>1.50</td>
<td>Norm</td>
<td>0.95</td>
<td>0.94</td>
<td>0.95</td>
<td>0.16</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Hetero</td>
<td>0.95</td>
<td>0.96</td>
<td>0.95</td>
<td>0.15</td>
<td>0.73</td>
</tr>
<tr>
<td>1.75</td>
<td>Norm</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>0.14</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Hetero</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.14</td>
<td>0.70</td>
</tr>
<tr>
<td>2.00</td>
<td>Norm</td>
<td>0.95</td>
<td>0.94</td>
<td>0.96</td>
<td>0.15</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Hetero</td>
<td>0.95</td>
<td>0.95</td>
<td>0.96</td>
<td>0.16</td>
<td>0.65</td>
</tr>
</tbody>
</table>

normal random effects, the standardized bias was 12% for the DPAM-Norm model and 14% for the DPAM-Het model. With random effects generated from a mixture distribution or a log normal distribution, the standardized bias, respectively, was $-1.02\%$ and $-0.82\%$ for the DPAM-Norm model and $-1.82\%$ and $-1.90\%$ for the DPAM-Het model. Root mean square error was similar, and in some cases improved, under data generated in violation of the normality assumption when compared to data generated in conformity with that assumption. Generally, RMSE deteriorated as treatment effect size and as covariate/confounder effect size increased, but it remained at acceptable levels.

Overall, the DPAM-Norm model was robust to the violations of normality simulated in this study, and the DPAM-Het model did not provide any substantial improvement over the normal model, and in some case, underperformed the former model.
4.5 Is the Ordinal Mixed Effects Model Robust to Violations of the Normality Assumption?

It will be recalled that the mixed-effects ordinal propensity score that is used in the dynamic propensity adjustment approach for longitudinal observational data is derived from a mixed-effects ordinal logistic regression model that assumes normality of the random effects distribution. The mixed effects propensity score is obtained by evaluating the estimated regression equation, which is a function of the maximum likelihood estimate of the fixed effects, $\hat{\beta}$, and the empirical Bayes estimates of the random effects. Most studies of the issue agree that estimates of fixed effects in generalized linear mixed models are generally robust to departures from the normality assumption but can be adversely affected when the true random effects distribution is highly skewed or bimodal. Almost all studies agree that misspecification of the random effects distribution adversely affects inference regarding the random effects.
As noted in Chapter 2, all of the studies evaluating the impact of random effects misspecification on estimation in generalized linear mixed models deal with the situation of binary response variables only. None, so far as we are aware, have evaluated directly the sensitivity of ordinal mixed effects models to distributional assumptions regarding the random effects. In order to investigate further the simulation results summarized above, as well as to investigate the more general question regarding the robustness of ordinal mixed effects models to violations of the normality assumption, we undertook a further simulation study.

4.5.1 Simulation Design

Data were simulated in the following manner: Similar to the preceding simulation study, a balance longitudinal study was simulated with 250 subjects and 8 observations per subject. A vector of predictors was generated for each subject at each treatment occasion ($X_1$ through $X_4$), two continuous ($X_1$ and $X_3$), two binary ($X_2$ and $X_4$), two constant across treatment
occasions ($X_1$ and $X_2$), two time-varying ($X_3$ and $X_4$). A correlation was imposed on the set of predictors and set to 0.20. An ordinal outcome was generated for each subject at each time point as follows: For each subject a random intercept was generated based upon three distributions: (1) normal, (2) a mixture of two normal distributions, and (3) zero-mean shifted log-normal. The association between predictors of the ordinal outcome was varied, with a log odds ratio of 1.25 and 2.00. A latent, continuous outcome was generated using a linear mixed effects model:

$$
\lambda_{ij} = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \nu_i + \varepsilon_{ij},
$$

(4.4)

where $\varepsilon_{ij}$ is distributed as a standard logistic, with variance $\frac{\pi^2}{3}$, i.i.d., and $\sigma^2_\nu$ is set such that the intraclass correlation between repeated doses, $\rho_k$, is equal to 0.40. The observed ordinal outcome was generated using the threshold concept, with thresholds set to guarantee adequate representation of each outcome level.

Parameters were estimated using a mixed effects regression model for ordinal outcomes, which incorporated the proportional odds assumption (McCullagh, 1980). The regression procedure was implemented using the R (R Core Team, 2017) ordinal package (Christensen, 2015). Quintiles of the linear predictor were calculated and each observation was categorized according to its quantile membership, both according to the data generating process and the estimated linear predictor. The latter was computed from the fixed effects parameter estimates and the empirical Bayes estimates of the subject-specific random intercepts.

4.5.2 Evaluation of Performance

The performance of the ordinal mixed effects regression model under the various simulation scenarios was evaluated as follows: Parameter estimates for predictors were assessed for power, coverage, standardized bias (BS), percentage bias (%Bias), and root mean square error (RMSE).
Performance of the predictions (individual outcomes or quintiles) compared to the true outcomes were evaluated using the gamma (Γ) statistic, Kendall’s $\text{Tau-b} (T_b)$, and Somers’ $d$ statistic. Gamma is a nonparametric measure of the strength and direction of association that exists between cross-classified ordinal variables that is based upon the number of concordant and discordant pairs, where two pairs of observations are concordant if the observation ranking higher on one ordinal variable also ranks higher on the other. The range of Gamma is (-1,1). Ties (in this context, correct predictions) are not included in the calculation. Kendall’s $\text{Tau-b}$ is similar to Gamma, but with a correction for ties, and its range is also (-1,1). Finally, Somers’ $d$ is an asymmetric modification of Kendall’s $\text{Tau-b}$, in which one variable is regarded as independent and the other as dependent. The range of Somers’ $D$ is (-1,1) (Agresti, 1990, pp. 20–23, 28, 34). (In this case, the prediction is regarded as the dependent variable, and the true response is regarded as independent variable.) If the cross-classified ordinal variables are independent, all these measures tend toward zero.

4.5.3 Simulation Results

Results of the simulation study are summarized in Table IV through Table VI. Table IV contains the results for the estimation of the fixed effects parameters. In accordance with previous work involving generalized linear mixed models with binary outcomes – reviewed in Chapter 2 – the fixed effect parameter estimates are robust to misspecification of the random effects distribution. The bias, standardized bias, and percentage bias are minimal in all combinations of covariate effect size and random effect distribution. In addition, the observed coverage probabilities are quite close to the nominal coverage probabilities at all simulation factor levels. RMSE is low across all simulation combinations, and actually is lower for all parameter estimates when the random effects distribution is non-normal.
### TABLE IV

**FIXED EFFECTS**

Effect Size = $\log(1.25)$

<table>
<thead>
<tr>
<th>Parm.</th>
<th>Distribution</th>
<th>Coverage</th>
<th>Power</th>
<th>Bias</th>
<th>SB</th>
<th>% Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>Normal</td>
<td>0.9469</td>
<td>0.4324</td>
<td>0.0024</td>
<td>0.0188</td>
<td>1.0815</td>
<td>0.1286</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>0.9530</td>
<td>0.7590</td>
<td>-0.0030</td>
<td>-0.0367</td>
<td>-1.3494</td>
<td>0.0820</td>
</tr>
<tr>
<td></td>
<td>Log Normal</td>
<td>0.9560</td>
<td>0.8210</td>
<td>-0.0034</td>
<td>-0.0459</td>
<td>-1.5449</td>
<td>0.0751</td>
</tr>
<tr>
<td>X2</td>
<td>Normal</td>
<td>0.9409</td>
<td>0.1421</td>
<td>-0.0007</td>
<td>-0.0026</td>
<td>-0.3033</td>
<td>0.2612</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>0.9440</td>
<td>0.2590</td>
<td>-0.0024</td>
<td>-0.0147</td>
<td>-1.0907</td>
<td>0.1660</td>
</tr>
<tr>
<td></td>
<td>Log Normal</td>
<td>0.9420</td>
<td>0.3360</td>
<td>0.0009</td>
<td>0.0063</td>
<td>0.4245</td>
<td>0.1510</td>
</tr>
<tr>
<td>X3</td>
<td>Normal</td>
<td>0.9439</td>
<td>0.9680</td>
<td>0.0024</td>
<td>0.0402</td>
<td>1.0536</td>
<td>0.0585</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>0.9580</td>
<td>0.9900</td>
<td>-0.0024</td>
<td>-0.0463</td>
<td>-1.0761</td>
<td>0.0519</td>
</tr>
<tr>
<td></td>
<td>Log Normal</td>
<td>0.9480</td>
<td>0.9920</td>
<td>-0.0007</td>
<td>-0.0140</td>
<td>-0.3279</td>
<td>0.0523</td>
</tr>
<tr>
<td>X4</td>
<td>Normal</td>
<td>0.9630</td>
<td>0.4945</td>
<td>-0.0001</td>
<td>-0.0006</td>
<td>-0.0309</td>
<td>0.1089</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>0.9510</td>
<td>0.5390</td>
<td>-0.0049</td>
<td>-0.0469</td>
<td>-2.2041</td>
<td>0.1048</td>
</tr>
<tr>
<td></td>
<td>Log Normal</td>
<td>0.9720</td>
<td>0.5740</td>
<td>0.0028</td>
<td>0.0275</td>
<td>-1.2466</td>
<td>0.1013</td>
</tr>
</tbody>
</table>

Effect Size = $\log(2.00)$

<table>
<thead>
<tr>
<th>Parm.</th>
<th>Distribution</th>
<th>Coverage</th>
<th>Power</th>
<th>Bias</th>
<th>SB</th>
<th>% Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>Normal</td>
<td>0.9439</td>
<td>0.9990</td>
<td>-0.0021</td>
<td>-0.0158</td>
<td>-0.3035</td>
<td>0.1333</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>0.9490</td>
<td>1.0000</td>
<td>-0.0008</td>
<td>-0.0096</td>
<td>-0.1208</td>
<td>0.0872</td>
</tr>
<tr>
<td></td>
<td>Log Normal</td>
<td>0.9530</td>
<td>1.0000</td>
<td>0.0007</td>
<td>0.0087</td>
<td>0.1003</td>
<td>0.0803</td>
</tr>
<tr>
<td>X2</td>
<td>Normal</td>
<td>0.9489</td>
<td>0.7665</td>
<td>0.0069</td>
<td>0.0263</td>
<td>0.9921</td>
<td>0.2612</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>0.9490</td>
<td>0.9800</td>
<td>-0.0023</td>
<td>-0.0137</td>
<td>-0.3382</td>
<td>0.1716</td>
</tr>
<tr>
<td></td>
<td>Log Normal</td>
<td>0.9540</td>
<td>0.9940</td>
<td>-0.0036</td>
<td>-0.0229</td>
<td>-0.5235</td>
<td>0.1581</td>
</tr>
<tr>
<td>X3</td>
<td>Normal</td>
<td>0.9569</td>
<td>1.0000</td>
<td>0.0010</td>
<td>0.0161</td>
<td>0.1452</td>
<td>0.0626</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>0.9420</td>
<td>1.0000</td>
<td>0.0003</td>
<td>0.0059</td>
<td>0.0500</td>
<td>0.0586</td>
</tr>
<tr>
<td></td>
<td>Log Normal</td>
<td>0.9540</td>
<td>1.0000</td>
<td>-0.0029</td>
<td>-0.0526</td>
<td>-0.4232</td>
<td>0.0558</td>
</tr>
<tr>
<td>X4</td>
<td>Normal</td>
<td>0.9519</td>
<td>1.0000</td>
<td>0.0017</td>
<td>0.0149</td>
<td>0.2521</td>
<td>0.1169</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>0.9450</td>
<td>1.0000</td>
<td>-0.0053</td>
<td>-0.0468</td>
<td>-0.7696</td>
<td>0.1141</td>
</tr>
<tr>
<td></td>
<td>Log Normal</td>
<td>0.9530</td>
<td>1.0000</td>
<td>-0.0010</td>
<td>-0.0093</td>
<td>-0.1408</td>
<td>0.1045</td>
</tr>
</tbody>
</table>
The results regarding outcome prediction are summarized in Table V. The performance of the ordinal random intercept model deteriorates when the normality assumption is violated. All three measures show a substantial positive association between the true outcome and the predicted outcome when the random effects are normally distributed, but the association drops when the random effects are non-normal. Not surprising, the effect is more pronounced for the smaller covariate effect size. As the effect size increases, the contribution of the empirical Bayes estimate of the random effects to the linear predictor index diminishes.

Table VI summarizes the simulation results as they pertain to the quantiles as predicted from the estimated ordinal random intercept model. Similar to the predicted outcomes, the association of true outcome quantile with predicted quantile is close to unity when the random effects are normally distributed but declines when the random effects are non-normal. Again, the impact of the violation of the normality assumption lessens as the predictor effect size increases.

<table>
<thead>
<tr>
<th></th>
<th>Distribution</th>
<th>Somers’ D</th>
<th>Gamma</th>
<th>Tau-b</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Normal</strong></td>
<td>0.8010</td>
<td>0.9638</td>
<td>0.8096</td>
<td></td>
</tr>
<tr>
<td><strong>Mixture</strong></td>
<td>0.6010</td>
<td>0.8896</td>
<td>0.6561</td>
<td></td>
</tr>
<tr>
<td><strong>Log Normal</strong></td>
<td>0.5784</td>
<td>0.8875</td>
<td>0.6440</td>
<td></td>
</tr>
<tr>
<td><strong>Effect Size = log(1.25)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Normal</strong></td>
<td>0.8387</td>
<td>0.9756</td>
<td>0.8429</td>
<td></td>
</tr>
<tr>
<td><strong>Mixture</strong></td>
<td>0.7969</td>
<td>0.9697</td>
<td>0.8104</td>
<td></td>
</tr>
<tr>
<td><strong>Log Normal</strong></td>
<td>0.7939</td>
<td>0.9705</td>
<td>0.8086</td>
<td></td>
</tr>
</tbody>
</table>

TABLE V
Caution dictates that no firm, general conclusions be drawn from a limited simulation study such as the one described here. However, these simulations support the conclusion that the fixed effects parameter estimates from random intercept ordinal regression models are robust to misspecification of the random effects distribution. Inference regarding these parameters may be adversely impacted from many factors, including model misspecification, but departures from the assumption of normality, upon which these models are based, do not appear to a source of bias or inefficiency. However, an analyst should have less confidence in the empirical Bayes estimates of the random effects if he or she suspects that the data violate the normality assumption. If there is reason to suspect that assumption, and research interest is focused on prediction of the random intercepts, then more robust methods, such as the heterogeneity model, should be employed.

This small simulation study also sheds some light on the relative performance of the DPAM-Het estimation approach versus the DPAM-Norm approach in the preceding simulation study. The stratification process in the treatment effectiveness model is based solely on the rank order of the estimated propensity scores, or more precisely, the quantile membership of the estimated
linear combination of predictors and predicted random intercepts ("linear predictor index"). If the estimation process is able to preserve the rank order and thus the quantile membership of the true propensity scores in the estimated linear predictor index, then the estimation procedure should perform adequately. According to the results presented in Table VI there is a strong, positive correlation between predicted and true linear predictor index quantiles, even when the true random effects distribution is non-normal. Granted, in the case of the lower predictor effect size, there is a decrease in the positive association between true and predicted quantiles; however, the results from the simulation study presented earlier suggested that this decrease is not significant enough to significantly impact the performance of the estimation procedure.
ALTERNATIVES TO STRATIFICATION IN THE USE OF THE
ORDINAL DYNAMIC PROPENSITY SCORE

5.1 Introduction

As indicated earlier (see p. 7) stratification is not the only form of adjustment that employs the propensity score to reduce the effect of confounding on the estimation of treatment effect in observational studies. In their original series of papers in which the concept and application of the propensity score was introduced, Rosenbaum and Rubin and colleagues show that matching on the propensity score (Rosenbaum and Rubin, 1985; Rosenbaum, 1989; Rubin and Thomas, 1996) and covariate adjustment using the propensity score (Rosenbaum and Rubin, 1983) can be employed to address confounding. These uses of the propensity score have been further developed, and the range of its uses has vastly expanded, since its introduction more than thirty years ago (Austin, 2011; Schafer and Kang, 2008).

The ordinal dynamic propensity score, as introduced by Leon and collaborators (2001), employs the subclassification strategy to adjust for potential confounding in longitudinal data with ordinal treatment levels. The dynamic propensity score for ordinal treatments has not been deployed as a matching variable, as a weighting variable, or as a covariate directly inserted into a linear mixed effects model. Therefore, the question arises whether the dynamic propensity score can be so employed, and to what degree these employments are effective in reducing bias introduced by confounding and whether they provide efficient estimates of treatment effect. Therefore, implementation of these alternative strategies for use of the dynamic propensity score were developed and their performance explored. The goal of this study was to
evaluate and compare the performance of the dynamic propensity score for ordinal dose using (1) subclassification on the dynamic propensity score, (2) matching on the dynamic propensity score, (3) weighting using the dynamic propensity score, and (4) covariate adjustment using the dynamic propensity score. In addition, these methods were compared with the performance of a model that does not adjust for the propensity score (the unadjusted model).

Stratification on the dynamic propensity score for ordinal dose has already been developed and has been extensively discussed in this paper. However, we describe the development and implementation of the other methods below. First, in order to introduce these alternatives, we need first to define a common target estimand. In the following we will estimate the relative effect of treatment level $D = d$ on outcome versus effect of reference treatment level $D = d'$, which we will denote as $\pi_{dd'}$.

5.2 Alternatives to Subclassification on the Propensity Score

5.2.1 Linear Mixed Effects Regression

In the unadjusted model, $\pi_{dd'}$ is estimated by the regression coefficients of a mixed effects regression model for continuous outcomes:

$$Y_{ij} = \beta_D D_{ij} + X_{ij} \beta_X + \varepsilon_i + \varepsilon_{ij},$$

where $D_{ij}$ is a vector of dummy variables for treatment level, $\beta_D$ and $\beta_X$ are vectors of regression parameters for the treatment levels and the covariates, respectively, $\varepsilon_i$ is the random intercept and $\varepsilon_{ij}$ are normally distributed random errors. No adjustment is made for the propensity score. The estimand $\pi_{dd'}$ is estimated by:

$$\pi_{dd'} = \mathbb{E}[Y(d)] - \mathbb{E}[Y(d')] = \beta_d - \beta_{d'},$$
where $\beta_d$ is the fixed effect regression parameter associated with the dummy variable indicating that treatment level $d$, and $\beta_{d'}$ is analogously defined.

5.2.2 Inverse Probability Weighting

Several alternative propensity-score-weighted estimators for $\pi_{dd'}$ have been proposed ( Hirano and Imbens, 2001; Lunceford and Davidian, 2004; Lee et al., 2010). We will employ a modification of a weighted propensity score estimator for multilevel data proposed by Li et al. (2013). They employ their weighted estimator in the context of clustered observational data where the treatment is binary. Our modification of the estimator makes it appropriate for an ordinal treatment and thus we can employ the ordinal dynamic propensity score of Leon and Hedeker (2005). Li et al. (2013) propose several alternative propensity score models and estimators of treatment effect. We will employ a modification of the nonparametric clustered estimator, combining it with a mixed effects ordinal model for the propensity score.

The estimation procedure proceeds as follows:

1. Estimate the propensity score for each observation using a mixed effects ordinal regression model for the propensity score, yielding a vector of propensity scores for each observation: $((\hat{e}_{ij}(d_1, x), \hat{e}_{ij}(d_2, x), \ldots, \hat{e}_{ij}(d_M, x)), \text{ for every } i \text{ and every } j$. In contrast with subclassification on the propensity score, the procedure uses the actual estimated propensity score, and not just the linear predictor.

2. Estimate $\pi_{dd',i}$ for each subject:

$$
\hat{\pi}_{dd',i} = \sum_{j=1}^{n_i} \frac{[I_{ij}(d)] Y_{ij} w_{ij}}{w_i^d} - \sum_{j=1}^{n_i} \frac{[I_{ij}(d')] Y_{ij} w_{ij}}{w_i^{d'}} ,
$$

where

$$
w_i^* = \sum_j [I_{ij}(d)] w_{ij},
$$
\[
w_{ij} = \frac{1}{[I_{ij}(d)] \hat{e}(d, x) + [I_{ij}(d') \hat{e}(d', x)]}.
\]

3. Take the mean weighted by the total weights in each cluster:

\[
\hat{\pi}_{dd'} = \frac{\sum_i w_i \hat{\pi}_{dd',i}}{\sum_i w_i}
\]

where

\[
w_i = \sum_{j=1}^{n_i} w_{ij},
\]

where \( \hat{\pi}_{dd'} \) is the estimated effect of receiving treatment \( d \) versus receiving treatment \( d' \). Note that \( w_{i1} \) is undefined if subject \( i \) does not receive dose \( d \) at any time \( j \); similarly for \( w_{i0} \). Such subjects must be excluded from analysis, potentially leading to inefficiencies if many subjects have little variation in dose levels over time. In addition, if enough subjects are excluded the estimand may change from the overall \( \pi_{dd'} \) to \( \pi_{dd'} \) for only those subjects who receive both treatments.

Standard errors of \( \hat{\pi}_{dd'} \) are obtained via parametric bootstrap sampling using subject-level resampling.

### 5.2.3 Matching on the Propensity Score

The purpose of matching in an observational study is to remove the influence of potential confounders by matching treated and control subjects on potential measured confounders so that, within each matched pair (or group), the subjects are as identical as possible with regarding to those confounders. Any difference in outcome can then be attributed to the treatment status of the respective subject, assuming no unmeasured confounding. In that sense matching
in an observational study attempts to replicate a block, randomized experiment. Clearly, this is an ideal only, since perfect matching is rare.

One difficulty with the matching strategy is that as the number of measured confounders grows, the likelihood of a good match on the high dimensional set of confounders decreases. Matching on the propensity score was introduced as a solution to this problem, as the propensity score is a scalar quantity. Matching on the propensity score also tends to balance the distribution of the measured confounders on which the propensity score is based.

Several different methods of matching have been proposed (Rosenbaum, 1989; Austin, 2014). Optimal matching is a technique whereby the total distance, however defined, between all possible pairs is minimized. It is often contrasted with greedy matching algorithms, such as nearest available matching, in which treated and controlled subjects are matched sequentially. Greedy matching algorithms generally do not minimized the total distance between pairs (Gu and Rosenbaum, 1993). Full matching is a technique in which the ratio of treated and control subjects within each matched stratum is allowed to vary, but each stratum contains either one treated subject and one or more controls or one control subject and one or more treated subjects. Rosenbaum has shown that the optimal match is always a full match (Rosenbaum, 1991).

Full matching is implement in several software packages, including the R packages, MatchIt (Ho et al., 2011) and OptMatch (Hansen and Klopfer, 2016). We utilize the OptMatch package, and the procedure for matching on the ordinal dynamic propensity score is as follows:

1. As in the case of subclassification on the dynamic propensity score, a ordinal random intercept model is fit, with treatment as the response variable and measured covariates as the predictors. Based upon the model fit, a linear predictor index ("the propensity score") is calculated for each observation.
2. Assuming the parameter of interest is $\pi_{dd'}$, the data set is winnowed to consist only of those observations in which either treatment $d$ or $d'$ is received. A matrix of propensity score discrepancies between all remaining observations is created.

3. A full matching is computed using the optimal matching algorithm.

4. An outcome contrast is computed for each matched stratum and an overall weighted average of these computed contrasts is computed. The stratum-specific estimator of the treatment effects is given by:

$$\Delta_s = \bar{y}_{sd} - \bar{y}_{sd'}, \quad s = 1, \ldots, S,$$

where $y_{sd}$ is the average response of the treated and $y_{sd'}$ is the average response among the untreated, respectively, in stratum $s$. $\pi_{dd'}$ is estimated by a weighted average of the stratum-specific estimators:

$$\tilde{\pi}_{dd'} = \sum_s w_s \Delta_s,$$

where $w_s$ are proportional to the stratum-wise harmonic means of treatment and control counts (Hansen and Klopfer, 2016).

5. Standard errors of $\tilde{\pi}_{dd'}$ are computed using bootstrap resampling with replacement at the subject level.

5.2.4 Covariate Adjustment using the Propensity Score

Finally, we directly insert the linear predictor estimate estimated from a random intercept regression model for the ordinal treatment into a linear random intercept regression model with no other covariates other than the treatment level indicators:

$$Y_{ij} = \beta_d d_{ij} + \beta_{e\tilde{\eta}}_{ij} + \nu_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, i = 1, \ldots, N,$$
where \( d_{ij} \) is a vector of dummy variables for treatment level, \( \beta_d \) is a vector of regression parameters for the treatment levels, \( \hat{\eta}_{ij} \) is the linear predictor estimate and \( \beta_e \) is its parameter, \( \nu_i \) is a random intercept for subject \( i \) and \( \epsilon_{ij} \) are normally distributed random errors.

The effect of treatment \( d \) versus treatment \( d' \) is estimated by:

\[
\hat{\pi}_{dd'} = \beta_d - \beta_{d'},
\]

where \( \beta_d \) is the regression parameter associated with the dummy variable indicating that treatment level \( d \), and similarly for \( \beta_{d'} \).

### 5.3 Simulation Design

The data generation scheme for this simulation study was similar to the one outline in Chapter 4. The predictor variables are generated in the same manner as in Chapter 4. Two different distributions for \( \nu_i \) were used to generate random intercepts: (1) a normal distribution, \( \nu_i \sim N(0, \sigma^2_\nu) \), and (2) a zero-mean shifted log-normal distribution.

The effect sizes for each of the four predictors of dose in the propensity model were varied (1.25, 2.0).

A latent, continuous dose was generated using a linear mixed effects model with a random intercept and standard logistic error terms. The intraclass correlation between repeated doses was set to 0.40.

The association of each covariate \((x_1 - x_4)\) on both the continuous effectiveness outcome and dose were specified to be approximately equivalent. The observed time-varying ordinal dose, \( T_{ij} \), was then generated using the threshold concept. A continuous outcome was generated using a linear random intercept regression model that included the effect of all the predictors as well as that of the treatment levels with independent and normally distributed random intercepts.
and random errors. The variance of the error terms and the random intercept were chosen such that the ICC for repeated outcomes was again 0.40.

A sample size of 250 subjects was generated for each simulated data set and each subject has 8 repeated observations. One thousand data sets were generated for each combination of simulation specifications. Each data set was analyzed separately using the adjustment strategies outlined above: the unadjusted model, propensity score stratification, weighting with the propensity score, matching on the propensity score, and direct adjustment for the propensity score in the treatment model.

We departed from the previous simulations by introducing misspecification in the models for the propensity score. Although all four covariates (time-varying and constant, binary and continuous) were used in the data generating process, only two of the four variables appear in the propensity models and the treatment effectiveness models. The misspecified propensity model and misspecified treatment effectiveness model only included two predictors: one time-invariant, dichotomous variable ($X_2$) and one time-varying, continuous variable ($X_3$). In this way, two confounds, $X_1$ and $X_4$, each components of the full propensity score, were ignored in the estimation of the propensity score and in the estimation of treatment effectiveness.

The performance of the estimation procedures were evaluated by computing the coverage probabilities, power, bias, standardized bias, percentage bias, and root mean squared error of the estimated average treatment effect of treatment level 3 versus treatment level 2. The true average treatment effect, according to the data generating process, was set at 0.21.

5.4 Simulation Results

The results of the simulations using normally distributed random effects are presented in Table VII. All estimation procedures perform reasonably well when the effect size is relatively small, even the unadjusted method. Coverage rates are close to nominal and bias is acceptable,
especially given that these estimates are based upon grossly misspecified models where two
of the four predictors in the data-generating process are excluded. RMSE is similar across
methods when the effect size is small.

<table>
<thead>
<tr>
<th>Method</th>
<th>est</th>
<th>cover</th>
<th>power</th>
<th>bias</th>
<th>sb</th>
<th>percb</th>
<th>rmse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect Size = 1.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.218</td>
<td>0.946</td>
<td>0.556</td>
<td>0.008</td>
<td>7.627</td>
<td>3.743</td>
<td>0.103</td>
</tr>
<tr>
<td>Stratify</td>
<td>0.211</td>
<td>0.946</td>
<td>0.482</td>
<td>0.001</td>
<td>1.300</td>
<td>0.679</td>
<td>0.110</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.221</td>
<td>0.932</td>
<td>0.476</td>
<td>0.011</td>
<td>8.746</td>
<td>5.004</td>
<td>0.121</td>
</tr>
<tr>
<td>Matching</td>
<td>0.223</td>
<td>0.940</td>
<td>0.365</td>
<td>0.013</td>
<td>9.071</td>
<td>6.135</td>
<td>0.143</td>
</tr>
<tr>
<td>Cov Adjust</td>
<td>0.217</td>
<td>0.950</td>
<td>0.549</td>
<td>0.007</td>
<td>6.438</td>
<td>3.190</td>
<td>0.104</td>
</tr>
<tr>
<td>Effect Size = 2.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.269</td>
<td>0.925</td>
<td>0.721</td>
<td>0.059</td>
<td>58.247</td>
<td>28.143</td>
<td>0.117</td>
</tr>
<tr>
<td>Stratify</td>
<td>0.286</td>
<td>0.903</td>
<td>0.637</td>
<td>0.076</td>
<td>60.970</td>
<td>36.182</td>
<td>0.146</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.276</td>
<td>0.899</td>
<td>0.561</td>
<td>0.066</td>
<td>49.371</td>
<td>31.623</td>
<td>0.150</td>
</tr>
<tr>
<td>Matching</td>
<td>0.374</td>
<td>0.772</td>
<td>0.705</td>
<td>0.164</td>
<td>101.302</td>
<td>78.297</td>
<td>0.231</td>
</tr>
<tr>
<td>Cov Adjust</td>
<td>0.267</td>
<td>0.910</td>
<td>0.698</td>
<td>0.057</td>
<td>52.261</td>
<td>27.286</td>
<td>0.124</td>
</tr>
</tbody>
</table>

All procedures perform relatively poorly when the effect size is large. All methods produce
upwardly biased estimates of the treatment effect of treatment level 3 versus treatment level
2. The matching method perform particular poorly on this metric. The matching method
also has a greatly elevated root mean square error, which may be an effect of the large bias of
the estimator or may indicate loss of efficiency as well. Coverage probability for matching is
also well below nominal probabilities. When the effect size is relatively large, the unadjusted
method and the method using covariate adjustment with the propensity score perform the best
on most metrics, although the stratified approach and the weighted approach are close behind them in performance.

The results of the simulations using the log normal distribution are summarized in Table VIII. Again, all methods perform relatively well when the effect size is small on metrics of bias and RMSE. Coverage probabilities are close to nominal for all methods when the effect size is small, bias is minimal, and RMSE is comparable among all methods. Again, for the larger effect size the performance of all approaches deteriorates substantially, with relatively large bias. The matching method and the weighted method perform the worst, with very large upward biases, poor coverage probabilities, and elevated RMSE. The stratification approach performs better than the alternatives in the context of the larger effect size. Although it too is upwardly biased, the stratified estimator, relative to the alternatives, is accurate; its percentage bias is 15% versus 80% for the weighted approach and 105% for the matching approach. The coverage probability for the stratified estimator is close to nominal probability and the RMSE remains similar to that of the estimators under the circumstances of a small predictor effect size.

In summary, this simulation study demonstrates that the stratified approach is more consistently accurate and efficient under the conditions simulated in this study. In addition, this study demonstrates the potential deleterious effects of model misspecification on estimation for all the proposed methods, especially when the effect sizes of omitted covariates is large. Of course, the conditions simulated in this study are limited. It would be of interest to see how the estimation approaches perform when the effect size of the predictors is not extreme, and when the propensity and outcome models do not differ so dramatically from the data-generating processes. These issues are a worthy subject of further investigation.
<table>
<thead>
<tr>
<th>Method</th>
<th>Est</th>
<th>Coverage</th>
<th>Power</th>
<th>Bias</th>
<th>SB</th>
<th>% Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Effect Size = 1.25</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.213</td>
<td>0.951</td>
<td>0.551</td>
<td>0.003</td>
<td>2.496</td>
<td>1.196</td>
<td>0.101</td>
</tr>
<tr>
<td>Stratified</td>
<td>0.205</td>
<td>0.952</td>
<td>0.523</td>
<td>-0.005</td>
<td>-4.360</td>
<td>-2.154</td>
<td>0.104</td>
</tr>
<tr>
<td>Matching</td>
<td>0.214</td>
<td>0.928</td>
<td>0.404</td>
<td>0.004</td>
<td>2.708</td>
<td>1.726</td>
<td>0.134</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.213</td>
<td>0.941</td>
<td>0.502</td>
<td>0.003</td>
<td>3.162</td>
<td>1.652</td>
<td>0.110</td>
</tr>
<tr>
<td>Cov Adjust</td>
<td>0.212</td>
<td>0.949</td>
<td>0.549</td>
<td>0.002</td>
<td>2.154</td>
<td>1.033</td>
<td>0.101</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Effect Size = 2.00</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.271</td>
<td>0.920</td>
<td>0.752</td>
<td>0.061</td>
<td>60.016</td>
<td>28.886</td>
<td>0.118</td>
</tr>
<tr>
<td>Stratified</td>
<td>0.240</td>
<td>0.944</td>
<td>0.623</td>
<td>0.030</td>
<td>28.968</td>
<td>14.476</td>
<td>0.109</td>
</tr>
<tr>
<td>Matching</td>
<td>0.432</td>
<td>0.628</td>
<td>0.866</td>
<td>0.222</td>
<td>150.479</td>
<td>105.841</td>
<td>0.267</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.277</td>
<td>0.904</td>
<td>0.616</td>
<td>0.067</td>
<td>52.901</td>
<td>31.816</td>
<td>0.143</td>
</tr>
<tr>
<td>Cov Adjust</td>
<td>0.271</td>
<td>0.919</td>
<td>0.749</td>
<td>0.061</td>
<td>59.664</td>
<td>28.855</td>
<td>0.118</td>
</tr>
</tbody>
</table>
CHAPTER 6

ANALYSIS OF NIMH COLLABORATIVE DEPRESSION STUDY

6.1 The CDS Data

The National Institute of Mental Health (NIMH) Collaborative Depression Study (CDS) is a longitudinal observational study of major affective disorders (major depressive disorder, mania, or schizoaffective disorder) conducted at five academic medical centers in the United States, with up to 20 years of follow-up data (Leon et al., 2003). The patient group that will be examined in our analysis will be taken from the 431 subjects who met the criteria for major depressive disorder at intake in the CDS and had no history of mania, hypomania, or schizoaffective disorder.

Given the variety of somatic treatments used to treat depression, which include a variety of medications and electroconvulsive therapy or ETC, a unipolar composite antidepressant rating has been developed as a summary measure of the intensity of somatic antidepressant treatment. This measure of somatic treatment intensity (\( tx \)) has five ordinal categories, ranging from 0 (no treatment) to 4 (intensive treatment). The temporal unit of analysis is a treatment interval, defined as the interval beginning with a change in treatment intensity in a patient suffering a major depressive episode and ending with a subsequent change in treatment intensity or the end of follow-up, whichever is later.

The outcome variable in this analysis is the change from the beginning to the end of the treatment interval on the Psychiatric Status Rating (PSR), a standardized rating scale used to measure the degree of mental illness (\( psr\_change \)). The PSR ranges from 1 (no symptoms) to 6 (definite criteria of major depression with prominent psychotic symptoms or extreme
impairment). Therefore, \texttt{psr\_change} is an integer variable that ranges from -5 to 5. However, for the purposes of this analysis, it will be treated as a continuous variable. (It should be noted that \texttt{psr\_change} is measured as the PSR score at the beginning of the treatment interval minus the PSR score at the end of the interval. Thus a reduction in depressive symptoms is associated with a positive \texttt{psr\_change}).

The analysis includes 2049 observations of 182 subjects who did not develop bipolar disorder during follow-up, recovered from the intake episode of major depression, had at least one subsequent depressive episode, had at least one change in treatment while in an episode subsequent to the intake episode, and had no missing data (Leon and Hedeker, 2007a).

This data set is the same as the one analyzed in papers by Leon and Hedeker (2003, 2007a). However, missing data were treated differently so results of our analysis may differ from that of Leon and Hedeker.

We describe our analysis of this data set using propensity adjustment models using stratification on the estimated dynamic propensity score, as this approach proved the most reliable of the methods considered in Chapter 5. In addition, we fit models for propensity for treatment intensity assuming that the random effects were generated from a mixture distribution, with increasing numbers of mixture components. By our model fit criteria, the model with a single component, i.e., the model described below, was the best.

The covariates entered into the model for propensity for treatment intensity and for treatment effectiveness were: (1) symptom severity over the 8 weeks prior to the treatment interval (\texttt{max8psr}), an ordered factor with six levels, (2) positive symptom trajectory over the past 8 weeks (\texttt{pos8traj}), a binary variable, (3) negative symptom trajectory over the past 8 weeks (\texttt{neg8traj}), also a binary variable, (4) number of months prior to treatment interval of the
depressive episode (\text{duration}), (5) age treated as a categorical variable (\text{age}), and (6) study site (\text{site}).

The symptom severity (\text{max8psr}) ranged from 1 to 6, with a mean of 4.6. 33\% of observations had a worsening symptom trajectory (\text{traj}), 51\% had a unchanged symptoms over the treatment interval, and 16\% showed improvement. The average number of months prior to the current treatment interval (\text{duration}) was 4.3. The distribution of the age of patients when observations were made, and the number of patients and observations at each study site are given in Table IX.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>CDS STUDY COVARIATES</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 30</td>
<td>14.6%</td>
</tr>
<tr>
<td>30-39</td>
<td>26.5%</td>
</tr>
<tr>
<td>40-49</td>
<td>21.7%</td>
</tr>
<tr>
<td>50-59</td>
<td>11.9%</td>
</tr>
<tr>
<td>60+</td>
<td>25.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study Site</th>
<th>Observations</th>
<th>Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>235 (11%)</td>
<td>18 (10%)</td>
</tr>
<tr>
<td>3</td>
<td>518 (24%)</td>
<td>58 (32%)</td>
</tr>
<tr>
<td>4</td>
<td>200 (9%)</td>
<td>21 (12%)</td>
</tr>
<tr>
<td>6</td>
<td>860 (39%)</td>
<td>53 (29%)</td>
</tr>
<tr>
<td>7</td>
<td>372 (17%)</td>
<td>32 (18%)</td>
</tr>
</tbody>
</table>

Figure 3 through Figure 7 provide a graphical representation of the conditional distribution of the predictors given treatment intensity level. A one-way analysis of variance (ANOVA) for episode duration by treatment intensity is significant at the 0.05 level for a difference in the
average episode duration among at least some of the levels of treatment intensity. Pearson’s chi-squared tests on the remaining categorical variables show an association with each of the predictor variables and treatment intensity, suggesting imbalance in those predictors over the treatment levels.

Figure 3. Mean Symptom Severity by Treatment Intensity

6.2 Results

6.2.1 Model for Propensity for Treatment Intensity

A propensity score was computed for each observation based on the estimated parameters from a ordinal random intercept logistic model assuming normality of the random effects, with the above identified covariates entered into the model. Observations were then stratified into
propensity score quintiles. Table XI is a cross classification table for treatment intensity versus propensity score quintile. There is good representation of each treatment intensity level in each propensity quintile, although, as expected, observations in lower propensity quintiles tend to receive lower treatment intensities, and observations in upper propensity quintiles tend to get more intense treatment.

6.2.2 Evaluation of Propensity-Adjusted Balance

The goal of propensity score adjustment is to obtain balance in the distribution of potential confounders. In order to test the degree of balance obtained after adjusting for the estimated dynamic propensity score, we fit separate mixed effects ordinal logistic models with treatment intensity as the outcome variable and propensity score quintile and each respective predictor as independent variables. These models were compared to simpler models with only the propensity
score quintile as an independent variable in the model. The results are summarized in Table XII. Symptom severity, duration of episode, symptom trajectory improving and symptom trajectory worsening all remained significantly associated with treatment intensity adjusting for propensity score quintile, suggesting that imbalance remained in the distribution of these covariates after adjustment for the propensity score quintile. Age at all levels was not significantly associated with treatment intensity after adjustment for the propensity score quintile. Study sites 3 and 6 remained significantly associated with treatment intensity after adjustment. One might hypothesize that the adjusted odds ratio for the predictors would be closer to unity in the propensity score adjusted model than in the unadjusted model (Leon and Hedeker, 2007a). However, this hypothesis was not born out here in every case. The estimated effect of symptom
severity, duration of episode, and both symptom trajectory variables actually increased, albeit slightly.

6.2.3 Model for Treatment Effectiveness

To test whether there is interaction between treatment effectiveness and propensity score quintiles, two ordinal mixed effects logistic models were fit: (1) a model with treatment intensity and all other predictor variables, and (2) a more general model with all terms in the preceding model and all treatment intensity by propensity score quintile interaction terms. Based upon the results of the respective model fits and the likelihood ratio test comparing the fit of the two models, we conclude that treatment by propensity quintile interaction is not significant (-2LL difference = 20.003, df = 16, p = .2201). As the interaction is deemed non-significant, the quintile-specific treatment effectiveness analyses were pooled, which yielded a pooled estimate
of the association between treatment intensity and change in symptom severity. The treatment intensity parameter estimates from the pooled analysis is shown in Table XIII. The effect of each treatment when compared to the effect of no treatment is statistically significant. The lowest treatment intensity level is associated with a 0.21 (95% CI: 0.11, 0.32) greater decrease in the PSR index when compared to no treatment. (Recall that \texttt{psr\_change} is measured as the PSR score at the beginning of the treatment interval minus the PSR score at the end of the interval. Thus a reduction in depressive symptoms is associated with a positive \texttt{psr\_change}) The most intense treatment is associated with a 0.37 point decrease in the PSR index relative to no treatment. The size of the effect increases monotonically with the dose, consistent with the hypothesis that more intensive treatment is associated with greater improvement in depressive symptomology.
### TABLE X

**MODEL FOR PROPENSITY FOR TREATMENT**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>CI Low</th>
<th>CI High</th>
<th>Z</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptom severity</td>
<td>1.15</td>
<td>1.05</td>
<td>1.27</td>
<td>3.17</td>
<td>0.002</td>
</tr>
<tr>
<td>Duration of episode</td>
<td>0.97</td>
<td>0.94</td>
<td>1.00</td>
<td>-2.05</td>
<td>0.040</td>
</tr>
<tr>
<td>Symptom trajectory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improving</td>
<td>1.09</td>
<td>0.86</td>
<td>1.38</td>
<td>0.68</td>
<td>0.497</td>
</tr>
<tr>
<td>Worsening</td>
<td>1.57</td>
<td>1.29</td>
<td>1.91</td>
<td>4.56</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30-39</td>
<td>1.74</td>
<td>1.29</td>
<td>2.37</td>
<td>3.60</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>40-49</td>
<td>1.87</td>
<td>1.32</td>
<td>2.64</td>
<td>3.51</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>50-59</td>
<td>1.29</td>
<td>0.86</td>
<td>1.93</td>
<td>1.21</td>
<td>0.225</td>
</tr>
<tr>
<td>60+</td>
<td>1.14</td>
<td>0.77</td>
<td>1.70</td>
<td>0.65</td>
<td>0.520</td>
</tr>
<tr>
<td>Study Site</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.35</td>
<td>0.21</td>
<td>0.58</td>
<td>-4.03</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>4</td>
<td>0.44</td>
<td>0.24</td>
<td>0.80</td>
<td>-2.68</td>
<td>0.007</td>
</tr>
<tr>
<td>6</td>
<td>0.66</td>
<td>0.41</td>
<td>1.08</td>
<td>-1.65</td>
<td>0.099</td>
</tr>
<tr>
<td>7</td>
<td>0.52</td>
<td>0.31</td>
<td>0.89</td>
<td>-2.41</td>
<td>0.016</td>
</tr>
</tbody>
</table>

### TABLE XI

**TREATMENT INTENSITY VS. PROPENSITY QUINTILE**

<table>
<thead>
<tr>
<th>Treatment Intensity</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>130</td>
<td>89</td>
<td>78</td>
<td>81</td>
<td>46</td>
<td>424</td>
</tr>
<tr>
<td>1</td>
<td>158</td>
<td>108</td>
<td>79</td>
<td>73</td>
<td>35</td>
<td>453</td>
</tr>
<tr>
<td>2</td>
<td>71</td>
<td>115</td>
<td>110</td>
<td>115</td>
<td>121</td>
<td>532</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>57</td>
<td>79</td>
<td>85</td>
<td>109</td>
<td>370</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>32</td>
<td>63</td>
<td>56</td>
<td>99</td>
<td>270</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>419</td>
<td>401</td>
<td>409</td>
<td>410</td>
<td>410</td>
<td>2049</td>
</tr>
</tbody>
</table>
### TABLE XII

**ADJUSTED AND UNADJUSTED PARAMETER ESTIMATES**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Unadjusted</th>
<th>Adjusted</th>
<th>OR</th>
<th>CI</th>
<th>OR_Q</th>
<th>CI</th>
<th>Pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptom severity</td>
<td>1.25</td>
<td>1.28</td>
<td>1.15, 1.36</td>
<td>1.18, 1.39</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration of episode</td>
<td>0.96</td>
<td>0.82</td>
<td>0.93, 0.99</td>
<td>0.79, 0.84</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symptom trajectory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improving</td>
<td>0.93</td>
<td>0.70</td>
<td>0.75, 1.16</td>
<td>0.56, 0.88</td>
<td>0.0019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worsening</td>
<td>1.67</td>
<td>1.79</td>
<td>1.40, 1.98</td>
<td>1.51, 2.12</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (vs. &lt;30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30-39</td>
<td>1.29</td>
<td>1.13</td>
<td>1.04, 1.61</td>
<td>0.93, 1.39</td>
<td>0.2235</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-49</td>
<td>1.16</td>
<td>0.99</td>
<td>0.92, 1.47</td>
<td>0.79, 1.23</td>
<td>0.9138</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50-59</td>
<td>0.84</td>
<td>0.91</td>
<td>0.62, 1.14</td>
<td>0.69, 1.20</td>
<td>0.5169</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60+</td>
<td>0.91</td>
<td>1.03</td>
<td>0.66, 1.25</td>
<td>0.81, 1.33</td>
<td>0.7975</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study Site (vs. 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.59</td>
<td>0.68</td>
<td>0.43, 0.80</td>
<td>0.54, 0.87</td>
<td>0.0020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.80</td>
<td>0.85</td>
<td>0.50, 1.28</td>
<td>0.60, 1.21</td>
<td>0.3613</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.46</td>
<td>1.30</td>
<td>1.09, 1.95</td>
<td>1.05, 1.60</td>
<td>0.0174</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.93</td>
<td>0.98</td>
<td>0.64, 1.33</td>
<td>0.74, 1.28</td>
<td>0.8588</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE XIII

**MODEL FOR TREATMENT EFFECTIVENESS**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>CI Low</th>
<th>CI High</th>
<th>Pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dose 1</td>
<td>0.21</td>
<td>0.11</td>
<td>0.32</td>
<td>0.0001</td>
</tr>
<tr>
<td>Dose 2</td>
<td>0.26</td>
<td>0.15</td>
<td>0.38</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Dose 3</td>
<td>0.35</td>
<td>0.21</td>
<td>0.49</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Dose 4</td>
<td>0.37</td>
<td>0.21</td>
<td>0.54</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
CHAPTER 7

CONCLUSION

The purpose of this study was to explore extensions of the adjustment strategy, first introduced by Leon et al. (2001), that employs a dynamic propensity score for ordinal treatment as a stratification variable to reduce bias from potential confounding. As the dynamic propensity score is estimated using a mixed effects regression model, and the estimation procedures for such models encode distributional assumptions regarding the random effects, the question arises whether the adjustment strategy is vulnerable to violations of this assumption and whether the procedure can be modified to loosen its reliance on these distributional assumptions. Hence, we adapted the heterogeneity model, which was developed by Verbeke and Lesaffre (1996), for use in estimation of the dynamic propensity score. The heterogeneity model relaxes the distributional assumptions regarding the random effects and uses a mixture of normals to approximate many distributions forms.

We conducted a Monte Carlo simulation study to determine if the heterogeneity model would perform better than the original model when the true data-generating distribution of the random effects was non-normal. The results of this study demonstrated that the heterogeneity model performed well when the random effects distributions was non-normal. However, the study also demonstrated that the dynamic propensity score adjustment strategy developed by Leon and colleagues (2001) is robust to violations of the distributional assumptions, at least under the scenarios simulated in this study. Given that the heterogeneity model is computationally intensive, and adds complexity to the model selection process, it is difficult to recommend it as a routine alternative to the original dynamic propensity score adjustment strategy.
Questions remaining in this area include whether other models that do not make the normality assumption, some of which were discussed in Chapter 2, would improve the estimation of the dynamic propensity score, especially under extreme departures from the normality assumption, such as random effects that follow a multi-modal degenerate distribution. Nonparametric approaches, such as tree-based estimation methods (Sela and Simonoff, 2012) or boosting methods (Zhu et al., 2015), which are less sensitive to parametric assumptions and were developed to improve predictive power, might be adapted for use in estimation of the dynamic propensity score, which is essentially a prediction problem.

We conducted a second simulation study to examine the question whether ordinal mixed effects regression model methods are sensitive to departures from distributional assumptions. We found that, like mixed effects regression models for binary outcomes, the estimation of fixed effects parameters were robust to violations of the distributional assumptions. However, the results were more equivocal with regard to the prediction of the random effects. The simulation results showed that the predictive power of the models deteriorated in scenarios where the data-generating distribution of the random effects was non-normal. Although the performance of the models to predict the quantile membership of the linear predictor index deteriorated in scenarios where the random effects were not normal, there remained a strong, positive association between true and predicted quantiles, providing some explanation of the results of the first simulation study. Further work in this area could include the evaluation of the performance of the ordinal mixed effects regression model under a greater variety of data-generating mechanisms and under more extreme departures from normality. In addition, the performance of mixed models with multiple random effects remains to be evaluated.

Finally, we explored alternatives to stratification on the estimated dynamic propensity score and proposed dynamic propensity score weighted estimators and estimators based on matching
on the estimated dynamic propensity score and employing the estimated dynamic propensity score as a covariate in the treatment effectiveness model. Our Monte Carlo simulation study evaluated the performance of these alternative methods relative to the stratification method under several distributional scenarios. This study also examined the performance of these models under the circumstances of model misspecification, in both the propensity model and the treatment effectiveness model. We found that performance of these alternative approaches was satisfactory when covariate and confounder effect sizes were relatively small, but the performance deteriorated for all methods when these effect sizes were larger. In addition, the stratification approach was less affected by the bias introduced by the omission of important covariates in the propensity and treatment effectiveness models.

There is much further work to be done in this area, however. Alternative weighting schemes in both the weighted estimator and in the weights applied to obtain an overall estimate in the matching approach could be explored. Other adjustment methods, such as regression estimation with propensity-related covariates, could be adapted for use with the dynamic propensity score (Schafer and Kang, 2008). Alternative matching strategies for multilevel data could be explored (Rickles and Seltzer, 2014; Steiner et al., 2013). Moreover, performance of methods based upon the dynamic propensity score could be compared with alternative non- or semi-parametric methods of dealing with time-dependent confounding, such as inverse probability weighted estimation of marginal structural models and g-estimation of structural nested models (Robins and Hernán, 2009).
CITED LITERATURE


NAME     Colin Hubbard

EDUCATION
B.A., Philosophy, Carleton College, Northfield, Minnesota, 1985
M.A., Philosophy, Loyola University Chicago, Chicago, Illinois, 1996
M.S., Biostatistics, University of Illinois at Chicago, Chicago, Illinois, 2002
Ph.D., Biostatistics, University of Illinois at Chicago, Chicago, Illinois, 2017

HONORS
Illinois Public Health Research Pre-Doctoral Fellowship, University of Illinois at Chicago
School of Public Health Traineeship Grants, University of Illinois Chicago

PUBLICATIONS