Kansas State University Libraries
New Prairie Press

Conference on Applied Statistics in Agriculture | 2012 - 24th Annual Conference Proceedings

TREATMENT HETEROGENEITY AND POTENTIAL OUTCOMES IN LINEAR MIXED EFFECTS MODELS

Troy E. Richardson

Gary L. Gadbury

Follow this and additional works at: http://newprairiepress.org/agstatconference


Part of the Agriculture Commons, and the Applied Statistics Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation

Richardson, Troy E. and Gadbury, Gary L. (2012). "TREATMENT HETEROGENEITY AND POTENTIAL OUTCOMES IN LINEAR MIXED EFFECTS MODELS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1037

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

TREATMENT HETEROGENEITY AND POTENTIAL OUTCOMES IN LINEAR MIXED EFFECTS MODELS

Troy E. Richardson and Gary L. Gadbury
Department of Statistics, Kansas State University, Manhattan, KS 66506

Abstract

Studies commonly focus on estimating a mean treatment effect in a population. However, in some applications the variability of treatment effects across individual units may help to characterize the overall effect of a treatment across the population. Consider a set of treatments, {T,C}, where T denotes some treatment that might be applied to an experimental unit and C denotes a control. For each of N experimental units, the duplet $\{r_{Ti}, r_{Ci}\}$, $i = 1,2,\dots,N$, represents the potential response of the i^{th} experimental unit *if* treatment were applied and the response of the experimental unit *if* control were applied, respectively. The causal effect of T compared to C is the difference between the two potential responses, $r_{Ti} - r_{Ci}$. Much work has been done to elucidate the statistical properties of a causal effect, given a set of particular assumptions. Gadbury and others have reported on this for some simple designs and primarily focused on finite population randomization based inference. When designs become more complicated, the randomization based approach becomes increasingly difficult.

Since linear mixed effects models are particularly useful for modeling data from complex designs, their role in modeling treatment heterogeneity is investigated. It is shown that an individual treatment effect can be conceptualized as a linear combination of fixed treatment effects and random effects. The random effects are assumed to have variance components specified in a mixed effects "potential outcomes" model when both potential outcomes, r_T, r_C, are variables in the model. The variance of the individual causal effect is used to quantify treatment heterogeneity. Post treatment assignment, however, only one of the two potential outcomes is observable for a unit. It is then shown that the variance component for treatment heterogeneity becomes non-estimable in an analysis of observed data. Furthermore, estimable variance components in the observed data model are demonstrated to arise from linear combinations of the non-estimable variance components in the potential outcomes model. Mixed effects models are considered in context of a particular design in an effort to illuminate the loss of information incurred when moving from a potential outcomes framework to an observed data analysis.

Key words: treatment heterogeneity, potential outcomes, subject-treatment interaction, mixed effects

1. Introduction

Treatment heterogeneity refers to the variability of a treatment effect across individuals in a population.The term *treatment effect* implies a comparison of one level of treatment against another. To state that a treatment effect varies across individuals implies that this comparison of treatment levels is made *within an individual*. Although, such variability has often been acknowledged as an important consideration in the application of experimental findings to prospective individual experimental units (EU), decisions about the use of treatment in EU's generally make use of statistical information gathered about the average or mean effect and then apply that same information to theindividual (cf. Marshall, 1997).It should be noted, however, that the estimated mean effect may be misleading when the effect of a treatment varies widely across individuals. If individual treatment variation is large with respect to the mean, then there may exist subpopulations in which a control produces a more favorable response compared with treatment even though the treatment appears to produce a more favorable response on average across the entire population. Standard analyses are unable to detect the existence of such subpopulations since individual treatment variability is confounded with experimental error in these standard designs. This paper explores issues that arise when estimating a variance of individual treatment effects. This variance serves to quantify the degree of treatment heterogeneity in a population. Results reported here should be useful for applications where estimatingthis variance, in addition to estimating a mean effect,may be of interest.

The analyses of many fundamental experimental designspreclude the identification and estimation of treatment heterogeneity.For those designs that permit a subject-by-treatment effect in the LM or LMM, a number of ways have been proposed to handle treatment heterogeneity. Wilk and Kempthorne (1955) modeled a subject-by-treatment effect as a fixed effect. First, they assumed a value of zero for the fixed subject-by-treatment effect in all subjects and all treatment combinations. Subsequent analyses assumed that the sum of fixed subject-by-treatment effects over all units in a population receiving a particular treatment combination was zero. Ghosh and Crosby (2005) utilized clustering techniques in a cross-over design to generate subgroups which they then considered replicates of one "subject" in order to estimate differences in subject-by-treatment effects. Kramer et al. (2011) presented a method in which they subtracted the estimated fixed effects from the observations in a cross-over design and applied principle component analysis to residualsso as to isolate a subject-by-treatment effect.

Gadbury and others (e.g., Gadbury and Iyer, 2000; Gadbury et al., 2001) defined a variance that quantified a degree of treatment heterogeneity, calling it subject-treatment (S-T) interaction, and then considered the issues involved when estimating this variance. Some of these results were summarized in Gadbury (2010). In these works, details were presented concerning a two-sample CRD with a covariate. They showed that the S-T variance is not directly estimable in most designs without assumptions, but bounds for it can be estimated. Many methods that estimate a variance associated with treatment heterogeneity are actually evaluating observable consequences of treatment heterogeneity (e.g., variability across subsets of a population). Other approachesmay make assumptions that are not verifiable in observable data. For example, one such assumption would be that an observable individual treatment effect in a cross-over design is equal to the true individual effect of treatment. The issues involved with making this type of assumption were recently discussed in Poulson et al. (2012).In Gadbury et al., (2003) a matched-pairs design was considered where outcomes were binary and in Albert et al. (2004) a blocked

design with binary outcomes was considered. The latter paper produced nonparametric estimates in a randomization based framework. For continuous outcomes, results for estimating individual treatment heterogeneity in designs beyond a two-sample CRD were derived in the context of finite population, randomization-based inference. This was done for a matched-pairs design and a balanced two-period-two treatment cross-over design (see Gadbury, 2010, for a summary of some results).

Randomization techniques for deriving estimators for an S-T variance become increasingly intractable as designs become more complex. This paper reconsiders results from Gadbury and others in the context of a linear mixed effects modeling framework. Such models are especially useful for modeling data from complex experiments. As such, their use for evaluating treatment heterogeneity seems especially attractive and will allow such evaluation in contexts far broader than those considered thus far.

Our approach here first considers a potential outcomes (Rubin 1974) analysis of data from a typical design using a LMMto help elucidate the role of treatment heterogeneity in a statistical analysis. In the following two sectionswe(i) develop a potential outcomes LMM-based approach, (ii) provide principles for relating the potential LMM to the "usual" observable LMM in a CRD and matched-pairs settings that can easily be extended to more complex situations, and (iii) present specific results in detail for the matched-pairs case. Then in Section 4, we quantify the relationship between the potential LMM and observable LMM using both a constructed data example and simulation,andthen discuss the required assumptions to equate the potential LMM and the observable LMM. A comparison of the two models quickly reveal components associated with treatment heterogeneity that are estimable in the potential LMMbut not in the observable LMM, at least not without non-trivial assumptions. Deriving the model for both observable data and potential outcomes data in any particular experimental design where treatment heterogeneity is of interest may help facilitatean understanding of the degree to which treatment heterogeneity can be evaluated in observable data.

2. Treatment Heterogeneity or S-T Interaction

Potential Outcomes

Consider a set of treatments, $\{T, C\}$, where T denotes some treatment that might be applied to an EU and C denotes a control that also might be applied to an EU. It is certainly plausible to extend these ideas to more than two levels of treatment, but for the purpose of this paper, we restrict ourselves to only T and C. For each EU, imagine the existence of a duplet $\{r_{Ti}, r_{Ci}\}$, which represents the potential response of the i^{th} EU*if* treatment were applied and the response of the i^{th} EU*if* control were applied, respectively. Notice that it is important to use terminology such as "imagine", "consider", or "conceptualize" when discussing potential outcomes as it is impossible to simultaneously observe all potential outcomes for a given experimental unit at a particular time. This constraint of a potential outcomes framework has been called the fundamental problem of causal inference. (Holland, 1986)

Though it is not possible to simultaneously observe both of these potential responses, the potential outcomes framework facilitates the definition of the *true causal effect*or*true individual difference*of T compared with$C$ of the i^{th} EU, denoted d_i,as

$$d_i = r_{Ti} - r_{Ci}. \quad (1)$$

New Prairie Press
http://newprairiepress.org/agstatconference/2012/proceedings/15

If d varies across EU's in a population—i.e. $var(d) > 0$—then treatment heterogeneity exists. It is the variance of these individual effects that quantifies the degree of S-T interaction. Note that this variance cannot be directly estimated using observable data because of the fundamental problem of causal inference.

Observable Outcomes and the Randomization Mechanism

As noted above, only one potential response may be observed for a given EU at a given time. We suppose random chance selects the observable responses from the potential responses.Define a random indicator variable,Z_i , such that

$$Z_i = \begin{cases} 1, if\ i^{th}\ experimental\ unit\ receives\ T \\ 0, if\ i^{th}\ experimental\ unit\ receives\ C \end{cases}$$

Define the observable outcome of the i^{th} experimental unit, R_i, as follows:

$$R_i = r_{Ti} \cdot Z_i + r_{Ci} \cdot (1 - Z_i)$$

where r_{T_i} and r_{C_i} are the potential responses of the i^{th} experimental unit. In potential outcomes literature, the probability distribution ofZ_i is referred to as the *randomization mechanism*.Once the samples have been selected, define the usual mean difference using the observable outcomes

$$\bar{D} = \bar{R}_{T\cdot} - \bar{R}_{C\cdot} = \frac{1}{n_T}\sum_{i=1}^{N} r_{Ti} \cdot Z_i - \frac{1}{n_C}\sum_{i=1}^{N} r_{Ci}\,(1 - Z_i)$$

where $\bar{R}_{T\cdot}$ is the arithmetic average of then_T responses for those units whose potential response under T was selected to be observed and $\bar{R}_{C\cdot}$ is the arithmetic average of then_C responses of those units whose potential response under C was selected to be observed. We distinguish$\bar{D}$from the true individual causal effect given in (1) by referring to$\bar{D}$as the *naïve difference* or the *naïve effect*. If, for example, $var(d) = 0$, then this naïve effect would be a good surrogate (and a good estimate with a random assignment mechanism) for a constant true effect.

Comparison of an individual quantity, like d_i, with a quantity summarizing a group of individuals, such as $\bar{D}$, may not be valid when individual effects vary. In some designs it may be possible to define related quantities to facilitate a reasonable comparison of the true causal effect and the naïve effect. For example, in a matched-pairsdesignas considered in this paper, the naïve paired difference in the i^{th}pair is

$$D_i = R_{Ti} - R_{Ci}.$$

In this case, D_i may be thought of as a naïve version of the true, individual causal effect for the two units in the i^{th} pair, which here would be given by $d_{i1}$and d_{i2}.

Statistical Propertiesof Potential Outcomes

In the potential outcomes framework, we conceptualize the experimental process as the selection of a finite set of duplets (*F*) from an infinite population of duplets (Ω). Each duplet contains the set of potential responses for an EU. A randomization mechanism is then employed to the duplets in *F*to select the observable response from the potential responses. As in the "usual" experimental setting, the end result is a collection of n_T EU's receiving T and n_C EU's receiving$C$.From an infinite population perspective, the duplets are independent of one another, and the potential responses within a duplet follow the following joint distribution:

$$\binom{r_{Ti}}{r_{Ci}} \sim \left\{ \binom{\mu_T}{\mu_C}, \begin{bmatrix} \sigma_T^2 & \rho \cdot \sigma_T \sigma_C \\ \rho \cdot \sigma_T \sigma_C & \sigma_C^2 \end{bmatrix} \right\} \tag{2}$$

It should be expected that the two potential responses are correlated as they are potential responses of the same individual under different treatment conditions. The correlation, however, is non-estimable due to the fundamental problem of causal inference.

Much work has been done to elucidate the statistical properties of d_i, defined in (1), under certain sets of assumptions. In particular, Neyman (1935) and Rubin (1974) demonstrated that assuming uniform randomization, the expectation of the naïve effect with respect to the randomization mechanism given the finite set *F*is the true mean causal effect. That is,

$$E_Z(\bar{D}|F) = \bar{d} = \frac{1}{N}\sum d_i$$

where $\bar{d}$ is the average true causal effect for all EU's in *F* (that is, a finite population mean treatment effect). Furthermore, it can be shown that

$$E_\Omega[\bar{D}] = E_\Omega[E_Z(\bar{D}|F)] = E_\Omega[\bar{d}] = \mu_d$$

where the unconditional expectation is with respect to the distribution in (2) from which the finite set *F* is selected, and where $\mu_d = \mu_T - \mu_C$.

Similarly,

$$var_\Omega[\bar{D}] = var_\Omega[E_Z(\bar{D}|F)] + E_\Omega[var_Z(\bar{D}|F)] = var_\Omega\left[\bar{d}\right] + E_\Omega[var_Z(\bar{D}|F)].$$

Notice that $var_\Omega[\bar{D}] \geq var_\Omega\left[\bar{d}\right]$ (Dawid, 2000) with equality iff $E_\Omega\left[var_Z\left(\bar{D}|F\right)\right] = 0$. The latter condition simply means that all of the variability in the estimator$\bar{D}$ for $\mu_d$is in the selection of the finite set *F* from the broader population.The inequality incorporates random variability resulting from the treatment assignment mechanism. More of this discussion can be found in Gadbury (2001).

3. Potential vs. Observable Linear Mixed Model (LMM)

Stroup (2011) developed a method termed What Would Fisher Do (WWFD) to correctly identify the components of the LMM. This method was based on the contribution Fisher made

to a discussion paper authored by Yates (1935) in which Fisher distinguishes between two aspects of an experiment, the topographical or design aspect and the treatment structure. Fisher noted that each aspect can be written down in such a way that the total degrees of freedom for the entire experiment are accounted for within each respective aspect. Fisher goes on to explain that that the choice of an experimental design could be regarded as the choice of which elements from the two aspects are selected to correspond. Consider the two-sample CRD in which no technical error is present and in which a random effect which arises from the distinct application of the j^{th} level of treatment to the i^{th} EU is permitted.To apply Stroup's WWFD method to the potential outcomes framework in a two-sample CRD, it may be helpful to consider the following plot plan:

EU	Part of Duplet Receiving T	Part of Duplet Receiving C
1	T	C
2	T	C
...	...	...
N-1	T	C
N	T	C

One can see that the potential outcomes framework for this design is constructed by conceptualizing two sets of responses, one set receiving T and the other set receiving C. Furthermore, each EU is represented in each set. The topographical structure for the potential outcomes framework and corresponding degrees of freedom can then be laid out as follows:

Topographical	
Source	d.f.
Set	2-1
EU	N-1
Set*EU	(2-1)*(N-1)
Total	2N-1

The analysis above was completely topographical. The treatment structure and its corresponding degrees of freedom can be laid out as follows:

Treatment	
Source	d.f.
Trt	2-1
Parallels	2*(N-1)
Total	2N-1

where "Parallels" represent the number of times a level of treatment must be prepared to accommodate a given sample size. In this case, there are two levels of treatment and each level of treatment must be prepared N times, therefore the degrees of freedom associated with

Parallels is 2*(N-1). Notice that the Topographical and Treatment aspects completely account for the total degrees of freedom in the experiment. To combine these two aspects, we choose the degrees of freedom associated with Trt in the Treatment table to correspond to the degrees of freedom associated with Set in the topographical table. Furthermore, we choose the degrees of freedom associated with Parallels in the Treatment table to correspond to the sum of the degrees of freedom associated with EU and Set*EU in the Topographical table. The resulting combined ANOVA table is given below by replacing "Set" with "Trt" everywhere "Set" appears in the Topographical table:

Topographical		Trt		Combined	
Source	d.f.	Source	d.f.	Source	d.f.
Set	2-1	Trt	2-1	Trt	2-1
EU	N-1	"parallels"	2(N-1)	EU	N-1
Set*EU	(2-1)*(N-1)			Trt*EU	(2-1)*(N-1)
Total	2N-1	Total	2N-1	Total	2N-1

Based on the combined ANOVA table above, the resulting potential LMM is

$$r_{ij} = \mu + s_i + \tau_j + s\tau_{ij}$$
$$i = 1,2,\dots,N\ subjects;\ j = T, C\ .$$

where s_i represents a random effect of the i^{th} EU, τ_j represents a fixed effect of the j^{th} level of treatment, and $s\tau_{ij}$ represents the random effect of the j^{th} level of treatment applied to the i^{th} EU. In a model assuming no technical error, $s\tau_{ij}$ should be considered the experimental error.

Invoking the randomization mechanism to produce and observable data set effectively removes one-half of the data, under uniform randomization. Again, it may be helpful to conceptualize the resulting observable data set with the following plot plan:

EU	Part of Duplet Receiving T	Part of Duplet Receiving C
1	T	~~C~~
2	~~T~~	C
…	…	…
N-1	~~T~~	C
N	T	~~C~~

Notice that each EU is now represented only once within a set instead of being represented in both sets so the "Set*EU" term is removed from the Topographical structure and replaced by an "EU(set)" term. Also notice that the degrees of freedom associated with Parallels is reduced since each level of treatment need be prepared only n times instead of N, where 2n=N. This alters the Topographical and Treatment structures as follows:

Topographical		Trt		Combined	
Source	d.f.	Source	d.f.	Source	d.f.
Set	2-1	Trt	2-1	Trt	2-1
EU(Set)	~~N-1~~ 2(n-1)	"parallels"	~~2(N-1)~~ 2(n-1)	EU(Trt)	~~N-1~~ 2(n-1)
Set*EU	(2-1)*(N-1)			Trt*EU	(2-1)*(N-1)
Total	~~2N-1~~ 2n-1	Total	~~2N-1~~ 2n-1	Total	~~2N-1~~ 2n-1

Based on this new Combined ANOVA table, the observable LMM can be written

$$R_{ij} = \mu + \tau_j + \varepsilon_{ij}$$
$$i = 1,2,\dots,n_j, \qquad j = T, C$$

where n_j is the number of EU's per level of treatment, such that $N = n_T + n_C = 2n$in a balanced, two-sampleCRD (i.e., $n_T = n_C = n$) and ε_{ij} is the "usual" error term in a two-sample CRD.

A direct relationship between the potential and observable models can be established by defining

$$\varepsilon_{ij} = s_i + s\tau_{ij}$$

Note that there is not enough experimental material in the observable model framework to estimate all effects of interestspecified in the potential model. In order to estimate a treatment effect in the observable model,only the linear combination of the variancecomponents of subject and subject-by-treatment effectscan be estimated. If the potential framework were feasible, both the variance of the subject effect and the variance of the subject-by-treatment effect would be estimable. Even for thissimple design, relating the quantities in an observable model to those in the potential model takes some thought. Still, it is necessary to highlight the information that gets lost as one moves from potential to observable data and, thus, what quantities in a model become non-estimable. The relationship between the potential model and observable model is not as explicit in more complicated designs.

Using Stroup's WWFD method, we adapted it to the potential outcomes framework and arrived at the following potential LMM in a matched-pairs design:

$$r_{ijk} = \mu + b_i + s_{j(i)} + \tau_k + b\tau_{ik} + s\tau_{j(i)k} \qquad (3)$$
$$i = 1,2,\dots,B\ pairs;\ j = 1,2\ subjects\ within\ a\ pair;\ k = T, C\ levels\ of\ treatment$$

whereb_i represents a random effect of the i^{th}pair (i.e.-block), $s_{j(i)}$represents a random effect of the j^{th} subject within the i^{th} block, τ_k represents a fixed effect of the k^{th} level of treatment, $b\tau_{ik}$ represents a random effect of the k^{th} level of treatment being applied to the i^{th} block and $s\tau_{jk(i)}$ represents a random effect of the k^{th} level of treatment being applied to the j^{th} subject

with the i^{th} block and should be considered experimental error.In a matched-pairs analysis, the $s\tau_{jk\,(i)}$ term represents a random subject-by-treatment or subject-by-control effect. As in Wilk and Kempthorne (1955), we assume no technical error.

Recall that observable data are conceptualized as being generated from potential outcomes by invoking a randomization mechanism that effectively removes one-half of the data. In this design, two of the four total potential outcomes within each pair are effectively removed so that there is one subject receiving treatment T and one receiving treatment C. By considering what information is "lost" by invoking the randomization mechanism, we use the potential LMM as a template to arrive at the observable LMM. This process is an important step in the appropriate estimation of effects in the observable model as misspecification of the model in SAS PROC GLIMMIX has been demonstrated to alter both model effect estimation and inference (Boykin et al., 2010). In this particular design, if each EU is permitted only one observable response instead of simultaneous potential responses, then we "lose" multiple observations per subject and a subject effect may no longer be estimated. Similarly, by invoking a randomization mechanism, only one level of each treatment is observable per pair and a block-by-treatment effect is no longer estimable. Thus by confounding these effects and defining the non-estimable portions of the potential LMM to be residual error, the resulting observable LMM in a matched-pairs design is

$$R_{ijk} = \mu + b_i + \tau_k + \varepsilon_{ijk}$$
$$i = 1,2,\dots,B\ pairs;\ j = 1;\ k = T, C\ levels\ of\ treatment$$

whereε_{ijk} is taken to be experimental error. The notation used here allows for straightforward extension to a block design with more than two subjects per block.

A direct relationship between the observable model and the potential model may be established by defining

$$\varepsilon_{ijk} = s_{j\,(i)} + b\tau_{ik} + s\tau_{j\,(i)\,k}$$

In a matched-pairs design, the ε_{ijk} term represents the random treatment error or the random control error.Under the assumption of unit-treatment additivity,$b\tau_{ik} = s\tau_{j\,(i)k} = 0\ for\ all\ i,j$, and

$$\varepsilon_{ijk} = s_{j\,(i)}$$

irrespective of the level of treatment assigned to the j^{th} EU. More discussion of this is in the next section.

4. An Illustration using Simulation

Illustration Using a Constructed Data Set

The results presented here are for a matched-pairs design, although we have extended the methods presented here to other designs as well. In order to demonstrate the utility of these techniques, we present here a constructed data example comparing the effects of two different types of laser surgery on visual acuity. This constructed data example is based loosely on the

analysis of an actual dataset (KARNS, 1993). The measure of visual acuity in the actual study was a count of correctly identified characters from a visual acuity chart. Here, the data were constructed to represent a change in the Logarithm of the Minimum Angle of Resolution (LogMAR) scores over a three-month period. Imagine that $N = 100$ EU's suffering from diabetic neuropathy were randomly assigned to receive red-krypton laser surgery in one eye and blue-green argon laser surgery in the other. Responses are the change in visual acuity measured from base-line to three months post-surgery. We simulated a potential dataset based on the potential model given in (3)whereb_i represents the random effect of the i^{th}EU, $s_{j\,(i)}$represents the random effect of the j^{th}eye within the i^{th}EU, τ_k represents a fixed effect of the k^{th} level of laser surgery with $T = Krypton$ and $C = Argon$, $b\tau_{ik}$ represents a random effect of the k^{th} level of laser surgery being applied to the i^{th}EU and $s\tau_{jk\,(i)}$ represents a random effect of the k^{th} level of laser surgery being applied to the j^{th}eye with the i^{th}EU. $s\tau_{jk\,(i)}$should be considered experimental error in the potential LMM. The distributional assumptions on random effects are as follows:

$$\begin{gathered} b_i \sim iid\ N(0, \sigma_b^2) \\ s_{j\,(i)} \sim iid\ N(0, \sigma_s^2) \\ b\tau_{ik} \sim iid\ N(0, \sigma_{bt}^2) \\ \begin{bmatrix} s\tau_{j\,(i)T} \\ s\tau_{j\,(i)C} \end{bmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{sT}^2 & 0 \\ 0 & \sigma_{sC}^2 \end{pmatrix}\right] \end{gathered} \qquad (4)$$

$b_i, s_{j\,(i)}$, $b\tau_{ik}$ and $s\tau_{j\,(i)k}$ are mutually independent.

Table 3(i) gives the values used in simulation.

Once a potential dataset had been constructed, an observable constructed dataset was produced by randomly selecting one potential response per eye within an EU to be the observable response for that eye. Table 1 gives the estimates and standard errors of identifiable quantities based on these two constructed datasets. Corresponding estimates of the same identifiable quantities in the potential and observable models should not be expected to be identical since the values of the observable estimates incorporate random variability resulting from the treatment assignment mechanism and will depend upon the particular realization of a vector of Z_i's.

The Quantity of Interest

Define the true causal effect for this experimental design, d_{ij}, to be the difference in the potential response of the j^{th} eye in the i^{th} EUundergoing laser surgery with Krypton and the potential response of the j^{th} eye in the i^{th} EUundergoinglaser surgery with Argon. From the linear model above,

$$d_{ij} = r_{ijT} - r_{ijC} = (\tau_T - \tau_C) + (b\tau_{iT} - b\tau_{iC}) + \left(s\tau_{j\,(i)\,T} - s\tau_{j\,(i)\,C}\right).$$

The EU effect and eye-within-EU effect are removed by virtue of the fact that under thepotential outcomes framework, both potential responses occur simultaneously in the same EU

Estimates: Potential Constructed Data		
Fixed Effects	**Estimate**	**Std. Error**
Argon	-0.1223	0.0159
Krypton	-0.0314	0.0154
Difference	0.0909	0.0158
Random Effects	**Var. Estimate**	**Std. Error**
EU	0.0112	0.0028
Eye(EU)	0.0017	0.0004
EU*Trt	0.0112	0.0018
Eye*Argon	0.0041	0.0007
Eye*Krypton	0.0012	0.0004

Estimates: Observable Constructed Data		
Fixed Effects	**Estimate**	**Std. Error**
Argon	-0.1229	0.0153
Krypton	-0.0378	0.0167
Difference	0.0851	0.0174
Random Effects	**Var. Estimate**	**Std. Error**
EU	0.0105	0.0028
Argon Error	0.0130	0.0030
Krypton Error	0.0173	0.0034

Table 1.*Estimates of Effect of Laser Therapy based on Constructed Datasets. Values represent estimates and standard errors of estimable quantities from the potential and observable constructed datasets.*

and the same eye-within-EU. Based on the model assumptions given in (4) and the simulation values in Table 3 (i), notice that

$$var(d_{ij}) = 2\sigma_{bt}^2 + (\sigma_{sT}^2 + \sigma_{sC}^2) = 0.030.$$

As noted in Table 1, the estimate of $var(d_{ij})$ from the constructed potential dataset is given by

$$\hat{var}(d_{ij}) = 2\hat{\sigma}_{bt}^2 + (\hat{\sigma}_{sT}^2 + \hat{\sigma}_{sC}^2) = 0.0276. \quad (5)$$

Contrast these results with that of the naïve difference for this experimental design, D_i, defined to be the difference in responses between the eye in the i^{th} EUactually assigned to receive laser surgery with Krypton and the eye in the i^{th} EUactually assigned to receive laser surgery with Argon. Since the naïve difference is defined to be across eyes, the eye-within-EU effect is not removed. That is

$$\begin{aligned} D_i = R_{ij} - R_{ij'} &= r_{ijT} - r_{ij'C} \\ &= (\tau_T - \tau_C) + (s_{j(i)} - s_{j'(i)}) + (b\tau_{iT} - b\tau_{iC}) + (s\tau_{j(i)T} - s\tau_{j'(i)C}). \end{aligned}$$

Based on the relevant model assumptions given in (4) and the simulation values given in Table 3 (i),

$$var(D_i) = 2\sigma_s^2 + 2\sigma_{bt}^2 + (\sigma_{sT}^2 + \sigma_{sC}^2) = 2\sigma_s^2 + var(d_{ij}) = 0.032 \quad (6)$$

Also notice that

$$var(D_i) \geq var(d_{ij}). \quad (7)$$

so that $var(D_i)$ is an estimable upper bound for $var(d_{ij})$. Using theobservable constructed dataset estimates in Table 1 to estimate $var(D_i)$ yields the following estimate:

$$\hat{var}(D_i) = \hat{\sigma}^2_{eT} + \hat{\sigma}^2_{eC} = 0.0313. \quad (8)$$

Notice that $\hat{var}(D_i) \geq \hat{var}\left(d_{ij}\right)$ and the estimates from the constructed datasets confirm the relationship between $var(D_i)$ and $var\left(d_{ij}\right)$ given in (7).

In addition to the estimate of σ_s^2, $\hat{\sigma}_s^2$, from the potential constructed dataset given in Table 1, it would seem reasonable to compute a second estimate of σ_s^2,$\tilde{\sigma}_s^2$, based on the relationship between $var(D_i)$ and $var\left(d_{ij}\right)$ given in (6). This second estimate is given as follows:

$$\tilde{\sigma}_s^2 = \frac{\hat{var}(D_i) - \hat{var}\left(d_{ij}\right)}{2} = 0.0019$$

Recall that the estimated value of σ_s^2 from the potential constructed dataset given in Table 1 is $\hat{\sigma}_s^2 = 0.0017$. The discrepancy between $\hat{\sigma}_s^2$ and $\tilde{\sigma}_s^2$ can be attributed to variability in the observable dataset resulting from invoking the randomization mechanism since the selection of different sets of potential responses as the observable responses will yield different values of $\hat{var}(D_i)$ and thus different values of $\tilde{\sigma}_s^2$.

Some general remarks about (7) are noteworthy regarding the matched-pairs experimental design. If the assumption of unit-treatment additivity holds, then neither a pair-by-treatment nor subject-by-treatment effect exist (i.e.—each effect is considered to be 0 with variance equal to 0). This implies that the variability of the true causal effect, d_{ij}, is 0, that is, it is a constant effect in the population. Thus any variability of the observable, naïve effect, D_i, is only a function of the variability due to subjects within a pair, σ_s^2. If the assumption of unit-treatment additivity does not hold, then the variability D_i may be thought of as a linear combination of the variability of subjects within a pair, σ_s^2, the variability arising from treatment being applied to a certain pair, σ_{bt}^2, and the variability arising from a treatment being applied to a subject within a pair, σ_{sT}^2 or σ_{sC}^2. Under the circumstances of perfect matching, (i.e.--$\sigma_s^2 = 0$), the variance of D_i is a linear combination of σ_{bt}^2, σ_{sT}^2, and σ_{sC}^2. It is under this circumstance (i.e., perfect matching) that the variances of the observable, naïve effect and the true causal effect are equal. Otherwise, $var(D_i) > var\left(d_{ij}\right)$. How well subjects are matched cannot be assessed in this design.

Simulation Method

We conclude by confirming the analytical results from the constructed data sets with simulations. Potential outcomes data were simulated assuminga matched-pairsdesign. A total of $S = 100$simulated datasetswere generated for each of the following numbers ofblocks of size $n = 2$: $B = 10, B = 30, and\ B = 100$. The resulting number of responses for one simulated dataset in the potential outcome framework is given by $2N = 2 \cdot Bn = 4B$ and the resulting number of EU's in one simulated observable experiment was given by $N = Bn = 2B$. SAS PROC GLIMMIX was then utilized on the simulated data to obtain REML estimates of: (i) the difference in fixed effects between the two potential outcomes within a subject, (ii) the variances

of random effects in the potential model, and (iii) the variance of the difference in the two potential outcomes, denoted $var(d)$.

Next, one-half of the data were removed to simulate observed data under uniformly random treatment assignment for a matched-pairs design. PROC GLIMMIX was again utilized on the observed data to obtain REML estimates of: (i) the difference in fixed effects between the two treatment groups, (ii) the variances of identifiable random effects in the observable model, (iii) the variance of the linear combination of non-identifiable random effects that constitute the residual term or error variance in the observable data model, and (iv) the variance of the paired difference in observable data, denoted $var(D)$.Then the empirical mean of the $S = 100$simulations was compared to the simulated value for each of the respective estimates. Simulation estimates were considered reasonable if the true simulated value fell with three (3) empirical sampling standard errors of the empirical mean of the $S = 100$ simulated datasets. Simulations were performed under the sameset of assumptions given in (4). As with the constructed datasets, Tables2 and 3 below give the values of the simulation parameters used in this study.

Simulation Results

Figure 1 illustrates the result in (7). The dotted lines represent the true value used in the simulation. The upper line corresponds to the simulated value of $var(D_i)$ and the lower line corresponds to the simulated value of $var(d_{ij})$. The difference between the upper and lower dotted line should be equal to $2\sigma_s^2$, as demonstrated in (6). Indeed, in these particular simulations, $\sigma_s^2 = 0.001$, thus the distance between the two dotted lines can be seen to be $2\sigma_s^2 = 2 \cdot 0.001 = 0.002$. Notice that when$B = 100$, the true simulatedvalue of $var(D_i)$ is within two empirical standard errors of the empirical mean of the $S = 100$ estimates. This would indicate that the REML estimates from the observable model are reasonable estimates of the sum of the variances of the confounded components from the potential model.

The results displayed in Figure 1 are typical of the results from these simulations. Tables 2 and 3give more specific results of all effects of interest based on $S = 100$ simulated data sets. Values represent the empirical mean and empirical sampling standard error of estimates across the $S = 100$ data sets. Table 2 gives results for the fixed treatment effect for the model fit to both potential and observable data, Table 3 shows the values used in simulation in the potential model and the results for the random effects in the observable model. In all cases, as the block sizeincreased from 10 to 30 to 100, the empirical sampling variability of the effect estimates around the true simulatedvalue decreased, as expected. For most effects under consideration, the true simulated value is within one or twoempirical standard errors of the empirical mean. True simulated values of all effects were within three empirical standard errors.

5. Summary

In this paper we showed that a linear mixed model applied to a potential outcome framework can be of pedagogical value in investigating estimability of treatment heterogeneity. By conceptualizing the true causal effect as a random variable with expectation $\tau_T - \tau_C$ and some finite variance, we permit the treatment effect to vary according to subject and estimate

thecomponent of the overall variability that is due to the subject-by-treatment effect. One benefit of

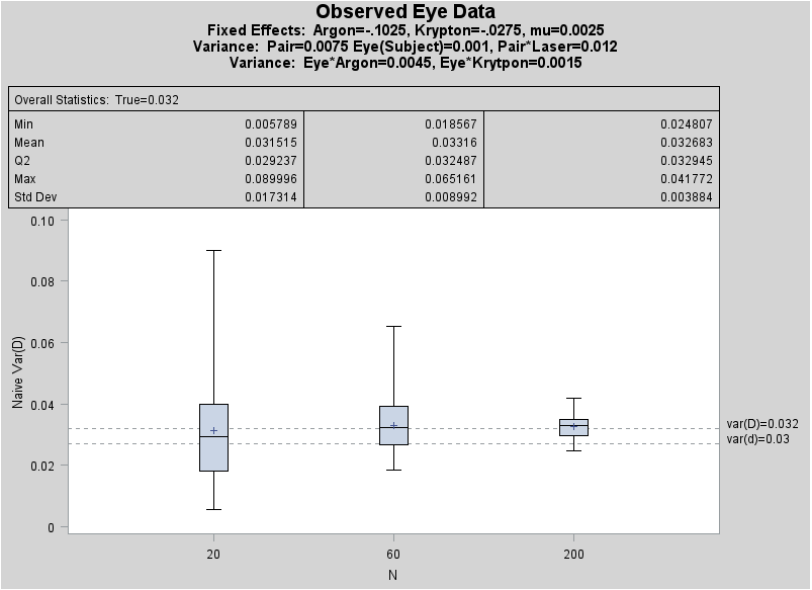

Figure1. *Empirical sampling distribution of estimated* $var(D_i)$. *Dotted lines represent values used in the simulation design.*

using potential outcomes to conceptualize this problem from a mixed model perspective is thatwe can clearly detail the "loss" of information that occurs when moving from a potential model to an observable data model. In a matched-pairs design, we described which effects were confounded when a treatment assignment mechanism is employed to generate observable data from potential outcomes. Furthermore, we demonstrated that the error effects in the observable data model are linear combinations of confounded effects from the potential model.

As one moves to a generalized block design, assigning more than one EU per block to receive each T and Cin the observable model facilitates the computation of more information about treatment heterogeneity within blocks. Cross-over designs that allow for "individual effects" to be observed provide information about individual treatment heterogeneity under different and perhaps more plausible assumptions from these other designs. Details about treatment heterogeneity in block designs and cross-over designs will be reported elsewhere.

In cases where treatment heterogeneity is suspected, it would be prudent to investigate this in addition to estimating a mean effect before a claim of the superiority of one treatment over another is established (Longford, 1999). LMM's are commonly used to estimate mean effects in various designs. As such, it is essentially "without cost" (in the sense that no new data are needed) to state the model that would be fit to potential outcomes data. A comparison between the two models delineates the information about causal effects that is lost in moving from potential to observable data, and what assumptions about non-estimable quantities (or design modifications) are needed to evaluate treatment heterogeneity in observable data.

Acknowledgements

We would like to thank the anonymous referee for critically reviewing the manuscript and offering helpful improvements, all of which we tried to incorporate.

Fixed Effect (Potential)	Simulated Value	2N	Average ($S = 100$)	Sampling Std. Error ($S = 100$)		Fixed Effect (Observable)	Simulated Value	N	Average ($S = 100$)	Sampling Std. Error ($S = 100$)
$\tau_T - \tau_C$	0.075	40	0.0709	.0061		$\tau_T - \tau_C$	0.075	20	0.0724	.0067
		120	0.0738	.0031				60	0.0738	.0033
		400	0.0762	.0018				200	0.0764	.0020

Table 2. *Fixed Treatment Effects. Values represent the average and empirical sampling standard error of treatment effect estimates across $S = 100$simulations in both the potential and observable data models for B=10, 30, and 100 blocks of size 2 EU's.*

Potential Model Component	Simulation Value
μ	0.0025
τ_T	-0.0275
τ_C	-0.1025
σ_b^2	0.0075
σ_{bt}^2	0.0120
σ_s^2	0.0010
σ_{sT}^2	0.0045
σ_{sC}^2	0.0015
$var(d_{ij}) = 2\sigma_{bt}^2 + \sigma_{sT}^2 + \sigma_{sC}^2$	0.0300

(i)

Observable Variance Component	Observable Simulation Value	N	Observable Average ($S = 100$)	Observable Sampling Std. Error ($S = 100$)
σ_b^2	0.0075	20	0.0082	.0007
		60	0.0070	.0005
		200	0.0073	.0002
Ctrl Error= $(\sigma_{bt}^2 + \sigma_s^2 + \sigma_{sC}^2)$	$(0.012 + 0.001 + 0.0015) = 0.0145$	20	0.0125	.0009
		60	0.0147	.0005
		200	0.0151	.0003
Trt Error= $(\sigma_{bt}^2 + \sigma_s^2 + \sigma_{sT}^2)$	$(0.012 + 0.001 + 0.0045) = 0.0175$	20	0.0174	.0012
		60	0.0184	.0007
		200	0.0176	.0003
$var(D_i) = 2\sigma_s^2 + var(d_{ij})$	$(2 \cdot 0.001) + 0.030 = 0.032$	20	0.0315	.0017
		60	0.0332	.0009
		200	0.0327	.0004

(ii)

Table 3.*Random Effects.(i) Values used in simulation for the potential model. (ii) Values represent the average and empirical sampling standard error of the variance estimates for random effects in the observable model across $S = 100$simulations for B=10, 30, and 100 blocks of size 2 EU's.*

References

Albert, J.M., Gadbury, G.L., and Mascha, E.J. (2005). Assessing Treatment Effect Heterogeneity in Clinical Trials with Blocked Binary Outcomes. *Biometrical Journal.* **47(5)**: 662-673.

Boykin, D., Camp, M.J., Johnson, L, Kramer, M., Meek, D., Palmquist, D., Vinyard, B., and West, M. (2010). Generalized Linear Mixed Model Estimation Using PROC GLIMMIX: Results from Simulations when the Data and Model Match, and when the Model is Misspecified. *In* Proceedings of the 22nd Annual Conference on Applied Statistics in Agriculture (*ed*. Weixing Song), Kansas State University, April 2010: 137-170.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*. **95**: 407 – 424.

Gadbury, G.L. (2010). Subject-treatment interaction. In *Encyclopedia of Biopharmaceutical Statistics, 3rd. Ed., Revised and Expanded.* Edited by Shein-Chung Chow. London: Informa Healthcare: 1316-1321.

Gadbury G. (2001). Randomization Inference and bias of standard errors. *American Statistician*. **55**: 310-313.

Gadbury, G.L., Iyer, H.K., and Allison, D. (2001). Evaluating subject-treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics.***11**: 313-333.

Gadbury, G.L., Iyer, H.K., Albert, J.M. (2004). *Journal of Statistical Planning and Inference*. **121**: 163-174.

Gadbury, G.L. and Iyer, H.K. (2000). Unit-Treatment Interaction and Its Practical Consequences. *Biometrics.***56**: 882-885.

Ghosh, S. and Crosby, H.R. (2005). Subject-treatment interactions in crossover trials: performance evaluation of subgrouping methods. *Journal of Statistical Planning and Inference*. **132**: 63-73.

Hinkelmann K. and Kempthorne O. *Design and Analysis of Experiments, Volume 1 :Introduction to Experimental Design*. Hoboken, NJ: John Wiley and Sons, Inc, 2008.

Holland, P.W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*. **81**: 945-960

The Krypton Argon Regression Neovascularization Study Research Group (KARNS) (1993).Randomized comparison of krypton versus argon scatter photocoagulation for diabetic neovascularization. *Ophthalmology***100**: 1655-1664.

Kramer, M., Chen, S.C., Gebauer, S.K., Baer, D.J. (2011). Estimating the subject by treatment interaction in non-replicated crossover diet studies. *In* Proceedings of the 23rd Annual Conference on Applied Statistics in Agriculture (*ed*. Weixin Yao), Kansas State University, April 2011: 96-110.

Longford, N. T. (1999). Selection bias and treatment heterogeneity in clinical trials. *Statistics in Medicine.* **18:** 1467 – 1474.

Marshall, A. (1997). Laying the foundations for personalized medicines. *Nature Biotechnology*. **15**: 954 – 957.

Neyman, J. (1935). Statistical Problems in Agricultural Experimentation (with discussion). *Supplement to the Jouarn1al of the Royal Statistical Society, Series B.* **2**: 107-180.

Poulson, R. S., Gadbury, G.L., and Allison, D.B. (2012). Treatment Heterogeneity and Individual Crossover Interaction. *American Statistician*. **66(1)**: 16-24.

Rubin, D.B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*.**66**:688-701.

Stroup, W.W. (2011). GLMM and "the Basics" – Paradigm Shift or Just My Imagination? *24th Annual Conference on Applied Statistics in Agriculture.* Manhattan, KS.

Wilk, M.B. and Kempthorne, O. (1955). Fixed, Mixed, and Random Models. *JASA*.**50**:1144-1167.

Yates, F. (1935). Complex Experiments. *Supplement to the Jouarn1al of the Royal Statistical Society.* **2(2)**: 181-247.