The Importance of Accounting for Correlated Observations

Kristin Sainani, PhD

INTRODUCTION

Improper analysis of correlated observations, such as repeated measurements on the same person, is a common error in medical studies. This article will review examples of correlated data, demonstrate the errors that arise when correlations are ignored, and discuss how to correctly analyze these data.

EXAMPLES OF CORRELATED OBSERVATIONS

Correlated data arise when pairs or clusters of observations are related and thus are more similar to each other than to other observations in the dataset. Observations may be related because they come from the same subject—for example, when subjects are measured at multiple time points (repeated measures) or when subjects contribute data on multiple body parts, such as both eyes, hands, arms, legs, or sides of the face. Observations from different subjects also may be related—for example, if the dataset contains siblings, twin pairs, husband-wife pairs, control subjects who have been matched to individual cases, or patients from the same physician practice, clinic, or hospital. Cluster randomized trials, which are performed to assign interventions to groups of people rather than to individual subjects (for example, schools, classrooms, cities, clinics, or communities), also are a source of correlated data because subjects within a cluster will likely have more similar outcomes than subjects in other clusters.

THE CONSEQUENCES OF IGNORING CORRELATIONS

Many statistical tests assume that observations are independent. The application of these tests to correlated observations will lead to the overestimation of *P* values in certain cases (when one considers within-subject or within-cluster effects) and underestimation in others (when one considers between-subject or between-cluster effects). These errors are illustrated in the following sections.

Within-Subject/Within-Cluster Comparisons

When subjects are compared with themselves under multiple treatments or at different time points, these are called within-subject comparisons; when they are compared with related subjects (such as twins), these are called within-pair or within-cluster comparisons. The advantage of comparing a subject to himself or herself or to a related person is that this comparison often results in considerable reduction in variability. Analyses that ignore the correlations will overestimate the variability, thus artificially increasing *P* values and decreasing the chances of observing a significant effect (decreasing the statistical power and increasing the type II error rate). Two examples follow that illustrate this problem.

Example 1. The authors of a recent randomized, blinded trial compared the efficacy of 2 sunscreens by using a split-face design [1]. Fifty-six subjects applied sunscreen with a sun protective factor (SPF) of 85 to one side of their face and an SPF of 50 to the other side of their face (the application sides were randomly chosen, and the sunscreen types were concealed) before spending 5 hours participating in outdoor sports on a sunny day. Investigators determined the occurrence of sunburn on each side of the participants' faces at

K.S. Department of Health Research and Policy, Stanford University, HRP Redwood Building, Stanford, CA 94305. Address correspondence to K.S.; e-mail: kcobb@stanford.edu
Disclosure: nothing to disclose

Disclosure Key can be found on the Table of Contents and at www.pmrjournal.org

PM&R
1934-1482/10/$36.00
Printed in U.S.A.

Vol. 2, 858-861, September 2010
DOI: 10.1016/j.pmrj.2010.07.482

Table 1a. *Original data table from Russak et al (1)*

Sun Protection Factor	Sunburned	Not Sunburned
85	1	55
50	8	48

$P = .03$, Fisher exact test.
Reprinted with permission [1].

Table 1b. *Correct presentation of the data from Russak et al (1)*

	SPF-50 Side	
SPF-85 Side	Sunburned	Not Sunburned
Sunburned	1	0
Not sunburned	7	48

$P = .0156$, McNemar exact test.
Reprinted with permission [1].

the end of the day. A person's tendency to burn on one side of his or her face is highly correlated with his or her tendency to burn on the other side. However, when the data were analyzed, these correlations were ignored: the authors reported that 1 of 56 participants were burned on the SPF 85 side of the face, whereas 8 of 56 were burned on the SPF 50 side ($P = .03$, Fisher exact test, Table 1a). This analysis treats all observations equally, as if there are 112 unrelated sides of the face. Table 1b shows the correct way to present and analyze the data.

Volunteers who burned on both sides of their face (n = 1) or neither side (n = 48) do not help us to discriminate between the performance of SPF 85 and SPF 50; only the volunteers who burned on a single side (n = 7) are informative. The correct analysis—called the McNemar exact test [2]—focuses only on these discordant subjects. In all 7 cases, the sunburn occurred on the SPF 50 side. The 2-sided *P* value associated with this extreme outcome (a 7-0 split) is .0156 (determined by a binomial distribution with n = 7 and $P = .5$). Thus the difference between the sunscreens is actually more significant than the authors have reported. Although the *P* values (.03 vs .0156) do not differ enough to change the study's conclusions, they can differ markedly in many cases, as the next example illustrates.

Example 2. Consider a simple hypothetical dataset in which investigators conducted a study with twins to examine the association of exercise with blood pressure. Six pairs of twins reported their physical activity levels and had their blood pressures measured. Investigators hypothesized that the more active twins would have lower blood pressures than the less active twins. The results are presented in Table 2.

The mean blood pressure for the more active twins is 3.5 mm Hg lower than for the less active twins (76.5 vs 80.0). If we ignore the correlations and analyze the data as 2 independent groups, this difference is not statistically significant ($P = .41$, 2-sample *t*-test). However, if we correctly analyze these data by focusing on the differences within twin pairs, it is statistically significant ($P = .02$, paired *t*-test). The *P* value is reduced because the variation in blood pressure within twin pairs (standard deviation = 2.6) is considerably less than between unrelated twins (standard deviation = 7.0 or 7.1) and because the paired *t*-test only has to account for one source of variability (variability within pairs) rather than 2 sources (variability from two groups of twins).

Between-Subjects/Between-Cluster Comparisons

When comparisons are made between unrelated subjects or clusters that have each received just one treatment, these are called between-subjects or between-cluster comparisons. In these situations, ignoring correlations in the data will lead to an underestimation of *P* values. For example, if a treatment works in a person's left eye, it is more likely to work in his or her right eye; thus it is unfair to count good outcomes in both eyes as 2 independent pieces of evidence for the treatment's effectiveness. Doing so artificially increases the sample size, decreases the *P* values, and potentially results in effects being

Table 2. *A simple hypothetical dataset involving correlated data (twin pairs)*

Twin Pair	Diastolic Blood Pressure in the Less Active Twin, mm Hg	Diastolic Blood Pressure in the More Active Twin, Mm Hg	Difference (More Active – Less Active), Mm Hg
1	87	82	−5
2	88	83	−5
3	80	78	−2
4	79	80	+1
5	77	71	−6
6	69	65	−4
Mean (SD)	80.0 (7.0)	76.5 (7.1)	−3.5 (2.6)
Test statistic	Two-sample *t*-test (incorrect analysis): $T_{10} = \frac{-3.5}{\sqrt{\frac{7.0^2}{6} + \frac{7.0^2}{6}}} = -0.86$, $p = .41$		Paired *t*-test (correct analysis): $T_5 = \frac{-3.5}{\sqrt{\frac{2.6^2}{6}}} = -3.31$, $p = .02$

Table 3. *A simple hypothetical dataset from a trial in which 50 subjects were randomly assigned to receive active drug (n =25) or placebo (n = 25) in both eyes*

Analysis	N (%) of Eyes Improving in the Control Group	N (%) of Eyes Improving in the Treatment Group	*P* Value	Odds Ratio and 95% Confidence Interval
Assuming eyes are independent*	17/50 (34)	27/50 (54)	.046	2.28 (1.02−5.11)
Correcting for within-subject correlation†	17/50 (34)	27/50 (54)	.11	2.28 (0.83−6.28)

*Data were analyzed with unconditional logistic regression.
†Data were analyzed by the use of a generalized estimating equation, correcting for within-subject correlation.

deemed significant when they should not be (a type I error). Two examples follow that illustrate this problem.

Example 1. In a hypothetical trial, 50 patients with bilateral eye disease were randomly assigned to receive an active drug or a placebo solution in both eyes (sample size per group is 25 patients [50 eyes]). Treatment was considered a success if symptoms improved by more than 50% in a given eye. Table 3 shows hypothetical results from this trial.

Strong agreement between eyes was found—80% of the subjects had the same outcome in both eyes (κ coefficient = .60). Thus treating the data as if there are 100 independent eyes will overstate the evidence for the drug's effectiveness. The informative sample size is actually somewhere between 100 and 50 (if there were perfect agreement between eyes, a subject's second eye would contribute no independent evidence of the drug's effectiveness and the sample size would be 50). The incorrect analysis (a χ^2 test or logistic regression) yields an artificially low *P* value of .046, whereas the correct analysis (a generalized estimating equation, corrected for within-subject correlation) yields a nonsignificant result of $P = .11$.

Example 2. Cluster-randomized trials are a common source of correlated data, but researchers often neglect the correlations in their analyses [3,4]. Calhoun et al [4] present a hypothetical example that shows the consequence of this failure. In this hypothetical randomized trial of an intervention to reduce physician error, 8 physicians were randomly assigned to a reduced shift length (n = 4) or control condition (n = 4). The outcome was the average number of charting errors per patient; data were obtained on 10 patients per physician for a total of 80 patients. Table 4 shows results from this hypothetical trial.

Observations made by the same physician will be highly correlated. For example, 2 of the 4 physicians in the intervention group are highly conscientious individuals who made no charting errors during the study period; thus it is clear that these 2 physicians each contribute just 1 unit of evidence for the intervention's effectiveness, not 10. If the data are analyzed as 80 independent observations (with use of a 2-sample *t*-test), the *P* value is highly significant, but the correct analysis (a hierarchical linear model) yields a nonsignificant result of $P = .273$.

HOW TO ADDRESS CORRELATED OBSERVATIONS

As the aforementioned examples demonstrate, correlated data require specialized statistical methods. Table 5 lists examples of statistical tests that assume independence and the corresponding tests for correlated data. For example, a 2-sample *t*-test is used to compare continuous, normally distributed outcomes between 2 independent groups, whereas a paired *t*-test is used to compare the same outcomes between 2 correlated groups.

Investigators often are less familiar with tests for correlated data than for uncorrelated data and may find them more challenging to implement and interpret. Thus many authors choose to handle correlations simply by removing them from the dataset. Although this approach is appropriate in certain situations, it often results in an unnecessary loss of information and statistical power. For example, one way to remove correlations is to change the unit of analysis. In the aforementioned physician study, the intervention was applied to physicians, not patients, so it makes sense to analyze the data at the physician rather than patient level; in this case, we would

Table 4. *A hypothetical cluster-randomized trial, from Calhoun et al (4)*

Analysis	Average Charting Errors From Control Physicians (n = 40 Patients, 4 Physicians)	Average Charting Errors From Treated Physicians (n = 40 Patients, 4 Physicians)	*P* Value
Assuming patients are independent*	2.75	1.7	<.0001
Correcting for within-physician correlation†	2.75	1.7	.273

*Data were analyzed with a 2-sample *t*-test.
†Data were analyzed by the use of hierarchical linear modeling.

Table 5. *Common statistical tests used to compare independent observations and the corresponding test for correlated observations, by the type of dependent (outcome) variable*

Dependent Variable	Test for Independent Observations	Corresponding Test for Correlated Observations
Continuous, normally distributed	Two-sample *t*-test	Paired *t*-test
Continuous or ordinal, non-normally distributed	Wilcoxon rank[illegible]st	Wilcoxon signed rank test
Continuous, normally distributed	ANOVA	Repeated-measures ANOVA
Continuous, normally distributed	Linear regression	Mixed models; hierarchical linear models
Binary/categorical	χ^2 test	McNemar χ^2 test (for 2×2 data)
Binary/categorical	Fisher exact test	McNemar exact test (for 2×2 data)
Binary/categorical	Logistic regression	Conditional logistic regression or generalized estimating equations

ANOVA = analysis of variance