Intraclass Correlation Coefficients Typical of Cluster-Randomized Studies: Estimates From the Robert Wood Johnson Prescription for Health Projects

David M. Thompson, PhD[1]
Douglas H. Fernald, MA[2,3]
James W. Mold, MD, MPH[4]

[1]Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma

[2]Department of Family Medicine, University of Colorado School of Medicine, Aurora, Colorado

[3]National Program Office, Prescription for Health, Aurora, Colorado

[4]Department of Family and Preventive Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma

MORE ONLINE
www.annfammed.org

ABSTRACT

PURPOSE Researchers who conduct cluster-randomized studies must account for clustering during study planning; failure to do so can result in insufficient study power. To plan adequately, investigators need accurate estimates of clustering in the form of intraclass correlation coefficients (ICCs).

METHODS We used data for 5,042 patients, from 61 practices in 8 practice-based research networks, obtained from the Prescription for Health program, sponsored by the Robert Wood Johnson Fund, to estimate ICCs for demographic and behavioral variables and for physician and practice characteristics. We used an approach similar to analysis of variance to calculate ICCs for binary variables and mixed models that directly estimated between- and within-cluster variances to calculate ICCs for continuous variables.

RESULTS ICCs indicating substantial within-practice clustering were calculated for age (ICC = 0.151), race (ICC = 0.265), and such behaviors as smoking (ICC = 0.118) and unhealthy diet (ICC = 0.206). Patients' intent-to-change behaviors related to smoking, diet, or exercise were less clustered (ICCs ≤0.007). Within-network ICCs were generally smaller, reflecting heterogeneity among practices within the same network. ICCs for practice-level measures indicated that practices within networks were relatively homogenous with respect to practice type (ICC = 0.29) and the use of electronic medical records (ICC = 0.23), but less homogenous with respect to size and rates of physician and staff turnover.

CONCLUSION ICCs for patient behaviors and intent to change those behaviors were generally less than 0.1. Though small, such ICCs are not trivial; if cluster sizes are large, even small levels of clustering that is unaccounted for reduces the statistical power of a cluster-randomized study.

Ann Fam Med 2012;10:235-240. doi:10.1370/afm.1347.

INTRODUCTION

Research conducted in practice-based research networks (PBRNs) often randomizes interventions, not by individuals, but according to a natural clustering unit, such as the physician or the practice in which patients receive care. Patients who receive care from the same physician or at the same practice are likely to resemble one another more than they do patients who see other physicians or attend other practices. These shared within-cluster characteristics may relate to patients inhabiting the same geographic area or community, or to commonalities in physician or practice styles. In studies involving multiple networks, clustering may also occur at the network level because of differences in geography or member selection. For example, some networks recruit mostly small

Conflicts of interest: authors report none.

CORRESPONDING AUTHOR

David M. Thompson, PhD
University of Oklahoma Health Sciences Center
Campus Mail CHB 321
PO Box 26901
Oklahoma City, OK, 73126-0901
dave-thompson@ouhsc.edu

rural practices, whereas others primarily include inner-city community health centers.

Researchers who conduct cluster-randomized studies must explicitly account for clustering at every stage of design and analysis. Failure to account for clustering, as well as the associations that naturally exist among patients within clinics, underestimates variability in outcome measures. Specifically, it underestimates the standard errors for between-subject effects, for example, effects related to cluster-randomized treatments. As a consequence, confidence intervals that estimate treatment effects will be too narrow and tests of hypotheses concerning treatment effects will be vulnerable to type 1 errors, the probability that investigators will declare differences to exist when they actually do not.[1]

The need to account for clustering begins during study planning, before data are collected or analyzed. Underestimating within-cluster variability in the study's primary outcome measures during design and planning will, in turn, underestimate the number of subjects necessary to detect a hypothesized treatment difference.[2] Failure to account for clustering can lead to studies that are underpowered.

To plan studies that have appropriate power, investigators need good estimates of clustering effects, typically in the form of intraclass correlation coefficients (ICCs). This coefficient, a parameter customarily signified as ρ, is defined as the proportion of a measure's total variance (σ^2_y) that is shared among members of defined clusters. Recognizing that an outcome's total variance (σ^2_y) is the sum of the between-cluster variance (σ^2_c) and the within-cluster variance (σ^2_w), then,

$$\rho = \sigma^2_c / \sigma^2_y = \sigma^2_c / (\sigma^2_c + \sigma^2_w).$$

If patients who attend the same clinic are relatively homogenous with respect to a measure, the within-cluster variance (σ^2_w) will be relatively small, and the between-cluster variability (σ^2_c) and ICC will be relatively large. When then between-cluster variability is large, it is difficult to attribute between-cluster differences to a treatment that is randomly assigned by cluster. As a result, studies that fail to account for this kind of clustering during their planning stage may be unable to detect treatment effects when they are executed.

Although investigators are most often interested in ICC estimates that quantify clustering in a study's outcome variable, ICCs are estimable for any variable measured in a sample. A population estimate for any variable's ICC is obtained using variance estimates for σ^2_c and σ^2_w that are derived from the sample. As are all population estimates, the ICCs are subject to some uncertainty, which is quantified in a confidence interval. Because an ICC's estimate involves a nonlinear combination of variances, the estimate's standard error and confidence interval involve calculations that are not straightforward. Obtaining confidence intervals for the ICC by bootstrapping[3] avoids this computational obstacle.

To plan cluster-randomized studies, investigators use the well-known variation inflation factor (VIF), generally expressed as $VIF = 1 + \rho(m-1)$, which requires estimates of the ICC (ρ) and of the study's mean cluster size (m). This formula for the VIF is based on a ratio that compares an outcome's variance in a study with independent clusters whose average size is m, with the outcome's variance calculated in a manner that ignores clustering and, instead, treats each patient as an independent cluster of size $m = 1$.[4] The Supplemental Appendix outlines the logic that underlies the formula, and is available at http://annfammed.org/content/10/3/235/suppl/DC1.

Also called the design effect, the VIF quantifies the effect that clustering among observations has on the variance of an outcome under study. Investigators use the VIF to produce both sample size calculations and hypothesis tests that are appropriately adjusted for the effect of clustering on an outcome's variance. Calculating a sample size that produces adequate power under the assumption that treatments are randomized at the level of the individual, but then multiplying that sample size by the VIF, ensures that a cluster-randomized design is of equal statistical power.[5(pp112-113)] Similarly, calculating χ^2 or t statistics to test hypotheses, while treating observations as unclustered, but then dividing these statistics by the VIF or the square root of the VIF, respectively, produces appropriate cluster-adjusted tests.[6(p333)]

METHODS

From 2003 to 2007, the Robert Wood Johnson Foundation (RWJF) funded 2 rounds of practice-based research network (PBRN) research on methods that might be used in primary care settings to identify and address 4 unhealthy behaviors: unhealthy eating, lack of physical activity, tobacco use, and alcohol overuse and abuse. Ten networks participated in the second round of the RWJF-sponsored Prescription for Health program and its Common Measures Better Outcomes (COMBO) study.[7-9]

One of the 10 networks enrolled only families with small children and another network enrolled only adolescents. Using data from the other 8 PBRNs, we calculated intraclass correlation coefficients (ICCs) for each of a list of patient-level behavioral and demographic variables and for certain physician and practice characteristics (Table 1). Table 1 organizes these

Table 1. Variables Analyzed

Variable	Level of Measurement	Description
Practice-level variables		
Number of physicians (FTE)	Continuous	
Number of staff (FTE)	Continuous	
Physician turnover rate	Continuous	Calculated as the number who left in past year, divided by total number
Staff turnover rate	Continuous	Calculated as the number who left in past year, divided by total number
Practice type	Binary	Solo or single specialty practice vs multispecialty practice
Use of electronic health record	Binary	Yes/no
Patient-level demographic variables		
Age	Continuous	Measured in years
Sex	Binary	Male vs female
Race	Binary	Nonwhite vs white
Patient-level measures of unhealthy behaviors		
Average number of drinks per day or month	Continuous	
Intention to reduce drinking alcohol	Continuous[a]	Applied only to patients who reported drinking at least 10 drinks in the last month
Smoking status	Binary	Smokers were identified as those who smoked at least part of a cigarette in the last 30 days
Intention to quit smoking	Continuous[a]	Applied only to patients who reported, on either the pre- or postintervention questionnaire, that they were former smokers, current smokers, or smokers trying to quit
Unhealthy diet	Binary	Unhealthy diet was defined as failure to consume 5 servings of fruit and vegetables a day.
Intent to improve diet	Continuous[a]	Applied only to patients whose responses indicated their diet was unhealthy
Physical inactivity	Binary	Physical inactivity was defined as no report of moderate or vigorous activity, nor of a 10-minute period of walking in the last 7 days
Minutes of physical activity on average day	Continuous	Calculated from reports of vigorous and moderate activity along with walking
Intent to start an exercise program	Continuous[a]	Applied to patients who reported less than 90 minutes of vigorous or moderate physical activity, including walking, in the last 7 days

FTE = full-time equivalent.

[a] Intention variables were measured on a 5-point ordinal scale, but were treated as continuous measures to calculate intraclass correlation coefficients.

variables and characteristics among 3 levels of the hierarchy within which observations were clustered: (1) patients within practices, (2) patients within PBRNs, and (3) practices within networks.

Patients Within Practices and Within Networks

The 8 PBRNs reported data on 5,042 patients who were aged at least 18 years and who received care in 61 practices. Networks that reported patient-level data included between 3 and 13 practices, and the practices enrolled between 1 and 364 patients. Although the 8 networks' projects differed in design, all collected practice-level data using the same practice information form and patient-level data on the same set of common measures.[8]

Practices Within Networks

While 61 practices contributed both patient-level and practice-level information, an additional 28 practices enrolled in the studies but contributed only practice-level data. Using data from these 89 practices, which numbered from 6 to 26 practices per network, we calculated ICCs on practice-level variables that included the number of full-time equivalent physician and staff, physician and staff turnover, and use of electronic medical records (Table 1).

Calculation of ICCs

The ICC is conventionally calculated using 2 quantities obtained from an analysis of variance.[10] One quantity is a mean square that estimates between-cluster variability (MSC), that portion of an outcome's variability that patients share because they are nested within clinics. The other quantity is a mean square that estimates within-subject variability (MSE) that is unique to (but assumed to be equal among) each subject regardless of cluster membership. These quantities are inserted into formulae established by Shrout and Fleiss,[10] the relevant one for this study being

$$\mathrm{ICC} = (\mathrm{MSC} - \mathrm{MSE})/(\mathrm{MSC} + (m - 1)\mathrm{MSE}).$$

Because the size of clusters typically varies in a cluster-randomized study, the formula for the ICC also

Table 2. ICCs (and 95% Empirical Bootstrap CIs), Adjusted Cluster Sizes (*m*), and VIFs Calculated for Patient-Level Variables (N = 5,042 Patients)

		Within 61 Practices			Within 8 PBRNs		
Variable	**n**	**ICC (95% CI)**	***m***	**VIF**	**ICC (95% CI)**	***m***	**VIF**
Age	4,984	0.151 (0.144-0.191)	80.08	12.94	0.054 (0.043-0.071)	554.35	30.63
Sex	5,004	0.050 (0.050-0 .089)	80.38	4.99	0.010 (0.006-0 .019)	555.93	6.68
Race	5,042	0.265 (0.246-0.296)	81.01	22.23	0.152 (0.133-0.175)	560.20	86.19
Smoking status	4,893	0.118 (0.117- 0.187)	78.60	10.19	0.072 (0.059-0.099)	543.43	40.15
Unhealthy diet	4,922	0.206 (0.178-0.252)	79.04	17.11	0.239 (0.197-0.284)	545.48	131.26
Inactivity	4,787	0.064 (0.062-0.095)	70.23	5.43	0.062 (0.054-0.092)	484.4	30.87
Minutes of physical activity per day	4,639	0.053 (0.051-0.094)	74.54	23.49	0.057 (0.046-0.082)	519.25	30.75
Average drinks per day	3,312	0.076 (0.067-0.142)	53.25	4.98	0.076 (0.054-0.111)	360.80	28.23
Average drinks per month	433	0.001 (0.000-0.103)	46.86	1.06	Data from 9 clinics within a single PBRN		
Intent to reduce drinking	193	0.207 (0.002-0.600)	18.18	4.56	Data from 9 clinics within a single PBRN		
Intent to quit smoking	378	0.000 (0.000-0.075)	40.74	1.00	Data from 9 clinics within a single PBRN		
Intent to improve diet	1,355	0.012 (0.003-0.037)	148.26	2.76	Data from 9 clinics within a single PBRN		
Intent to increase exercise	917	0.007 (0.000-0.042)	100.27	1.65	Data from 9 clinics within a single PBRN		

ICC= Intraclass correlation coefficient; PBRN = practice-based research network; VIF = variance inflation factor.

requires combining each cluster's size (m_k) to calculate an overall weighted mean cluster size (m).[6(equation 8)] This calculation of m was also necessary to VIFs.

$$m = \frac{1}{k-1}\left(n - \frac{\sum_k m^2_k}{n}\right)$$

We used an analysis of variance approach promoted by Reed[11] and Taljaard et al[12] to arrive at ICCs for binary variables. The approach is equivalent to that of a mixed model that estimates a random intercept for each cluster.

We arrived at ICCs for each continuous variable by directly estimating the between-cluster (σ^2_c) and within-cluster (σ^2_w) variances in a mixed model that treated clusters as random effects.[13(pp329-339)] The models, calculated in SAS PROC MIXED 9.2 (SAS Institute Inc), were structurally equivalent to hierarchical models where, for example, observations on patients were nested within either clinics or networks. These models estimated σ^2_c and σ^2_w using restricted maximum likelihood estimation, which produces more unbiased estimates than maximum likelihood estimation when observations are clustered or correlated.[14(p101)] We also used this mixed model approach to calculate ICCs for ordinal variables that reflected patients' intention to change health-related behaviors.

Point estimates for the ICCs are accompanied by 95% confidence intervals. To avoid the complicated estimate of a standard error that is required for an estimate that, similar to the ICC's, involves a nonlinear combination of variances, we calculated bootstrap 95% confidence intervals. Specifically, we resampled with replacement to produce 1,000 bootstrap samples,[3] calculated the ICC for each sample so obtained, then reported empirical 95% bootstrap confidence intervals. These intervals' limits are simply the ICC values that demarcate the 2.5th and 97.5th percentiles of the estimate's bootstrap distribution.

RESULTS

Patients Within Practices and Within Networks

Table 2 summarizes, for variables measured on individual patients, calculated ICCs and their 95% confidence intervals, adjusted cluster sizes (m), and VIFs. Large ICCs that reflect substantial clustering of patient characteristics within physician practices were evident for demographic such variables as age (ICC = 0.151) and the proportion of patients who are nonwhite and white (ICC = 0.265). Large ICCs were also found for such behaviors as smoking status (ICC = 0.118) and unhealthy diet (ICC = 0.206). The extent of within-patient clustering for alcohol use depended on how the behavior was measured; the ICC was estimated to be 0.076 when we assessed drinks per day but only 0.001 when we assessed average drinks per month. Relatively small ICCs (0.007 or lower) were calculated for the intention to change behaviors related to smoking, diet, and exercise. Patients' intent to change these behaviors was relatively diverse within the practices.

Corresponding ICCs within networks were generally smaller, which suggests that, even though within-practice clustering was evident for many measures, practices within the same network were relatively heterogenous with respect to the measures.

Table 3. ICCs (With 95% Empirical Bootstrap CIs) and VIFs Calculated for Practice-Level Variables, Collected on 89 Practices Within 8 PBRNs

	Within Network Statistics	
Practice Level Variable	**ICC (95% CI)**	**VIF**
Practice type	0.294 (0.187-0.499)	3.82
Use of electronic medical record	0.229 (0.101-0.406)	3.20
Number of physician FTEs	0.053 (0.000-0.427)	1.51
Number of staff FTEs	0.036 (0.000-0.407)	1.35
Number of staff/physician FTEs	0.062 (0.000-0.393)	1.60
Physician turnover rate	0.110 (0.029-0.564)	2.06
Staff turnover rate	0.066 (0.000-0.327)	1.63

FTE = Full-time equivalent; ICC = intraclass correlation coefficient; PBRN = practice-based research network; VIF = variance inflation factor.

Note: crude mean cluster size = 11.125; adjusted mean cluster size = 10.59 clinics per network.

Practices Within Networks

Table 3 summarizes the ICCs and VIFs calculated for practice-level variables, measured in 89 practices within 8 PBRNs. The project did not collect data at the level of individual physicians. Whereas practices within PBRNs were relatively homogenous with respect to practice type (ICC = 0.29) and the use of electronic medical records (ICC = 0.23), they were less homogenous with respect to their size and to the rate of turnover of physicians and staff.

DISCUSSION

Our analyses suggest that the ICCs for certain measures of health behavior are small, generally less than 0.1. Bland[4] describes this magnitude as typical for outcome variables in cluster-randomized studies. Though small, these ICCs are not trivial; if cluster sizes are large, even small levels of clustering, if unaccounted for, can reduce a study's statistical power.

We found larger ICCs for patient-level demographic variables and for practice-level variables, such as the presence of an electronic medical record, a measure that relates to the control of clinical processes. In this regard, the study reinforces others' observation that clustering is less evident for outcome variables than for other independent and process variables.[12]

High levels of within-practice clustering among demographic and other independent variables underscore the need, when analyzing data from studies that randomize interventions among practices, to adjust for confounding that arise as a result of between-cluster differences. Statistical methods exist to adjust for confounding. Moreover, where outcomes are measured on continuous scales, mixed or hierarchical models can adjust for practice- and patient-level clustering among covariates. For outcomes that are binomial or measured as counts, marginal models that use generalized estimating equations can derive cluster-adjusted estimates of treatment and other effects and are applicable as long as clusters are numerous.[15]

Because methods are available to adjust for clustering in the analysis of data that have already been collected, the primary use for information about clustering is for study planning. Investigators can use estimates of ICCs such as those provided here to ensure that a planned cluster-randomized study affords adequate power to detect a treatment effect. Investigators can initially use conventional sample size estimation techniques to determine that a sample of, for example, 100 independent and randomly selected subjects affords an 80 percent power to detect a prespecified and clinically meaningful effect. They can proceed to calculate a VIF by using a published estimate of the ICC for the study's outcome measure, along with an estimate of the study's likely cluster size. By multiplying the conventional sample size estimate by the VIF, the investigators can arrive at an appropriately inflated or augmented goal for subject recruitment. Recruiting a sample of this increased size ensures that, under the planned cluster randomization, the study affords 80% power to detect the prespecified effect.

This study provides estimates of the ICC at 3 levels of clustering: patients within practices, patients within networks, and practices within networks. The study estimated ICCs for binary outcome and process measures using an approach similar to analysis of variance (ANOVA) that, although advocated by Reed,[11] may apply only to data which, like the COMBO data, involved large clusters. Estimates of the variances that make up the ICC were robust in that we obtained similar results whether we used Reed's single-factor ANOVA approach[11] or the hierarchical models constructed using SAS PROC MIXED.[13] To estimate ICCs for binary and ordinal measures from studies with smaller clusters, a more appropriate approach might construct hierarchical logistic or cumulative logistic regression models, respectively, in software such as SAS PROC GENMOD, which can apply appropriate distributional assumptions along with generalized estimating equations methodology.[16,17]

In addition to estimating ICCs, this study provides confidence intervals on those estimates. Obtained by resampling, these intervals provide investigators with realistic ranges for the ICCs' true values. In particular, the intervals' upper bounds will generate the largest and most conservative VIFs that investigators might

use in calculating sample size estimates for cluster-randomized studies.

Investigators who plan studies with interventions that are randomized not to individuals, but to relatively homogenous groups or clusters of individuals, must account for clustering, particularly when planning the size of the studies' samples. A standard approach multiplies initial sample size estimates, made on the assumption that individuals are heterogenous and not clustered within groups, such as medical practices, by a variance inflation factor calculated on the basis of approximate cluster size and an estimate of the appropriate ICC. This study used data from the RWJF-sponsored Prescription for Health program, and its COMBO study[7-9] to provide point estimates and confidence intervals for ICCs for health behaviors and other patient- and practice-level characteristics. These estimates will be of interest to practice-based researchers as they plan research on similar health outcomes and patient behaviors.

To read or post commentaries in response to this article, see it online at http://www.annfammed.org/content/10/3/235.

Key words: intraclass correlation coefficient; primary care; cluster-randomized trial; practice-based research network; estimation techniques

Submitted February 18, 2011; submitted, revised, October 30, 2011; accepted October 13, 2011.

Funding support: This research was supported by The Robert Wood Johnson Foundation, Princeton, New Jersey.

Acknowledgments: We would like to thank the primary care practices around the country who participated in round 2 of the Prescription for Health studies.

References

1. Zyzanski SJ, Flocke SA, Dickinson LM. On the nature and analysis of clustered data. *Ann Fam Med*. 2004;2(3):199-200.
2. Guittet L, Giraudeau B, Ravaud P. A priori postulated and real power in cluster randomized trials: mind the gap. *BMC Med Res Method*. 2005; 5:25 http://www.biomedcentral.com/1471-2288/5/25. Accessed May 22, 2010.
3. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC; 1993.
4. Bland JM. Sample size in guidelines trials. *Fam Pract*. 2000;17(Suppl 1):S17-S20.
5. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. American ed. New York, NY: Oxford University Press; 2000.
6. Wears RL. Advanced statistics: Statistical methods for analyzing cluster and cluster- randomized data. *Acad Emerg Med*. 2002; 9:330-341.
7. Cifuentes M, Fernald DH, Green LA, et al. Prescription for Health: changing primary care practice to foster healthy behaviors. *Ann Fam Med*. 2005;3(Suppl 2):S4-S11.
8. Fernald DH, Froshaug DB, Dickinson LM, et al. Common measures, better outcomes (COMBO): a field test of brief health behavior measures in primary care. *Am J Prev Med*. 2008;35(5)(Suppl):S414-S422.
9. Balasubramanian BA, Cohen DJ, Clark EC, et al. Practice-level approaches for behavioral counseling and patient health behaviors. *Am J Prev Med*. 2008;35(5)(Suppl):S407-S413.
10. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428.
11. Reed JF III. Adjusted chi-square statistics: application to clustered binary data in primary care. *Ann Fam Med*. 2004;2(3):201-203.
12. Taljaard M, Donner A, Villar J, et al; World Health Organization 2005 Global Survey on Maternal and Perinatal Health Research Group. Intracluster correlation coefficients from the 2005 WHO Global Survey on Maternal and Perinatal Health: implications for implementation research. *Paediatr Perinat Epidemiol*. 2008;22(2): 117-125.
13. Singer JD. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *J Educ Behav Stat*. 1998;24(4):323-355.
14. Fitzmaurice G, Laird N, Ware J. Estimation and statistical inference. In: *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley; 2004: 87-102.
15. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health*. 2004;94(3):423-432.
16. Feng Z, Diehr P, Peterson A, McLerran D. Selected statistical issues in group randomized trials. *Annu Rev Public Health*. 2001;22:167-187.
17. Isaakidis P, Ioannidis JP. Evaluation of cluster randomized controlled trials in sub-Saharan Africa. *Am J Epidemiol*. 2003;158(9):921-926.