Design of Cluster-Randomized Trials of Quality Improvement Interventions Aimed at Medical Care Providers

Robert J. Glynn, PhD, ScD, M. Alan Brookhart, PhD, Margaret Stedman, MPH, Jerry Avorn, MD, and Daniel H. Solomon, MD

Background: Randomized trials aimed at improving the quality of medical care often randomize the provider. Such trials are frequently embedded in health care systems with available automated records, which can be used to enhance the design of the trial.

Methods: We consider how available information from automated records can address each of the following concerns in the design of a trial: whether to randomize individual providers or practices; clustering of outcomes among patients in the same practice and its impact on study size; expected heterogeneity in adherence and the response to the intervention; eligibility criteria and the trade-offs between generalizability and internal validity; and blocking or matching to alleviate covariate imbalance across practices.

Results: Investigators can use available information from an automated database to estimate the amount of clustering of patients within providers and practices, and these estimates can inform the decision on whether to randomize at the level of the patient, the provider, or the practice. We illustrate calculation of the anticipated design effect for a proposed cluster-randomized trial and its implications for sample size. With available claims data, investigators can apply focused eligibility criteria to exclude subjects and providers with expected low compliance or lower likelihood of benefit, although possibly at some loss of generalizability. Chance imbalances in covariates are more likely when randomization occurs at the level of the practice than at the level of the patient, so we propose a matching score to limit such imbalances by design.

Conclusions: Challenges to compliance, expected small effects, and covariate imbalances are particularly likely in cluster-randomized trials of quality improvement interventions. When such trials are embedded in medical systems with available automated records, use of these data can enhance the design of the trial.

Key Words: cluster randomized trials, design effect, contamination, quality improvement, intraclass correlation

(*Med Care* 2007;45: S38–S43)

Well-designed randomized trials provide the best evidence on the value of interventions to improve the quality of medical care. Although observational studies can yield useful insights on the value of practice patterns, concerns about selection of practices and subjects, unmeasured confounding variables, and concurrent changes in practice patterns and outcomes limit precise measurement of the impact of practice in observational studies.

A key issue in the design of trials to evaluate quality improvement interventions is the unit of randomization. This can be the individual patient, the provider, a group of providers within a practice, or several practices perhaps within a geographic area. If possible, randomization of individual patients has some key advantages because it allows for evaluation of treatment effects within practices. Randomization of individual patients also reduces the likelihood of covariate imbalances that can arise when practices see different types of patients. Furthermore, a trial that randomizes individual patients can use blocking within practices to limit the threat to validity that arises when practices have different levels of adherence to a protocol or when rates of outcomes vary across practices because providers' clinical skills differ. A trial that randomizes individual patients can also increase treatment effects, although at some loss of generalizability, through restriction to patients with documented eligibility and interest in the topic of the intervention.

In some situations, randomizing individual patients is either infeasible or unethical. If interventions are directed toward providers, they may find it impossible to deliver different, randomly assigned interventions to their different patients. If education or evidence leads providers to conclude that 1 therapeutic approach is preferred over others, they cannot ethically deliver an alternative. Furthermore, if patients share information with others treated by the same provider, assignment of all patients of this provider to the same intervention can enhance compliance, and conversely assignment to different interventions can lead to dilution of

From the Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, and Harvard Medical School, Boston, Massachusetts.

Supported by NIH grant AG18833.

Reprints: Robert J. Glynn, PhD, ScD, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont St., Suite 3030, Boston, MA 02120. E-mail: rglynn@rics.bwh.harvard.edu.

ISSN: 0025-7079/07/4500-0038

treatment effects through nonadherence. Contamination that occurs because patients do not adhere to their assigned treatment introduces a bias that well-designed studies seek to minimize. These design features often dictate that randomization be at the level of the provider.

Cluster randomized trials are more difficult to design than individually randomized trials for several reasons: the need to account for correlations in outcomes among patients within a practice, the particular ethical challenges in such trials, the likelihood that the unit of analysis differs from the unit of randomization, frequently limited data for eligibility criteria before randomization, and the heightened potential for imbalanced covariates.[1–3] However, randomized trials that evaluate interventions to improve the quality of health care are often conducted within health care systems in which administrative data are available to aid in the design of the trial.

This article considers how one can use such data to inform key aspects of study design. These elements include whether randomization should occur at the level of the practice or individual providers; what magnitude of clustering should be anticipated among patients treated by the same provider; what eligibility criteria would strengthen the trial; what covariates are likely to influence adherence to treatments and study outcomes in important ways; and how the design can best balance these covariates across randomized treatments.

UNIT OF RANDOMIZATION

A first question in the design of a trial to evaluate a quality improvement intervention is whether randomization should occur at the level of the patient, provider, practice, or even a broader group of related practices. If outcomes are measured separately in each patient, then randomization at the finest possible level has advantages for the balance of covariates. However, if providers within a practice share patients or treatment strategies, adherence to protocol-specified practices will probably decline. Thus, measures of agreement in treatments received by patients of different providers within a practice can index the likely contamination that would occur if these providers were randomized to different treatment groups.

In a placebo-controlled trial, nonadherence biases estimated treatment effects in an intention to treat analysis toward the null; it also reduces the power of the trial to find significant relationships. Specifically, the power of a study with p percentage of participants who comply with the assigned protocol, relative to a study with full compliance, is no more than p percent, and may be as low as p^2 percent, of the power of a fully compliant study.[4] For example, if a study is designed to have 90% power to detect an effect of a prespecified magnitude, but in the actual trial 20% of randomized participants cross over to the alternative treatment, then actual power will be no more than 72% (80% of 90%) and may be as low as 58% (64% of 90%). From the parallel perspective of the necessary sample size to detect a prespecified effect with given power,[5] if N is the number of subjects needed in each of 2 arms of a trial assuming perfect compliance, but a proportion p_1 of individuals assigned to active treatment do not comply and a proportion p_2 of individuals assigned to placebo independently initiate the active treatment, then the number needed per arm accounting for these drop-outs and drop-ins is $\mathrm{N}/(1 - p_1 - p_2)^2$ as long as p_1 and p_2 are not too large. Thus, even modest levels of contamination can have a large impact on the validity of a trial.

Consider, as an example, the choice of unit of randomization in a trial that evaluated a strategy to improve prescribing of antibiotics among inpatients in an academic medical center.[6] The outcome of interest, measured in individual patients, was an order for an antibiotic (levofloxacin or ceftazidime) deemed unnecessary by a priori criteria. Although individual interns, residents, and attending physicians wrote such medication orders, these doctors worked closely together within a service. The information from the educational intervention, directed to interns and residents and based on principles of academic detailing,[7,8] would likely have been shared among the physicians within a service. Because of this clear expectation of shared information, the units of randomization in the trial were 17 general medical, oncology, and cardiology services within the medical center. To provide some balance in patient covariates by design, randomization was blocked by service type (general medical, oncology, or cardiology).

In the design phase of some trials, the likely extent of nonadherence to the protocol that will arise through shared patients and practices is often unclear. Availability of a computerized claims database offers the potential to evaluate the extent to which multiple providers within a practice share patients and the extent of correlation of practice patterns among providers within a practice. One simple measure of sharing would be the percent of potentially eligible patients who see multiple providers within a practice. More sophisticated measures of clustering within practices could use multilevel models to quantify agreement in treatments among patients of the same provider, as well as among patients treated by different providers within the same practice.[9] Estimating the intracluster correlation among providers within a practice can inform the decision about the best level of randomization. A high intracluster correlation indicates that the practice would be preferred over individual clinicians as the more appropriate unit of randomization. Indeed, if patients or practice styles are shared among providers within a practice, then contamination would more likely occur if patients within the practice were assigned to different treatments.

In choosing the unit of randomization, a hybrid approach with physicians or practices randomized at 1 level, and patients separately randomized to a complementary intervention, may warrant consideration. For example, some practices might be randomly targeted for visits from academic detailers to discuss optimal prescribing for specific indications. A separate, independent randomization of at-risk patients from both intervention and control practices could choose some patients to receive additional information on optimal drug use either by mail, telephone, or both. Investigators could use available information in a database as they design the study to balance the practices on important covariates and to identify the at-risk patients. The design of the

Healthy Bones Project provides an example of such a trial.[10] This trial randomized 826 primary care physicians in Pennsylvania either to receive a one-on-one academic detailing encounter covering fall and fracture prevention and osteoporosis diagnosis and treatment, or to receive no intervention. The trial separately randomized, at the level of the patient, 31,715 of these physicians' patients who were enrolled in a state-run pharmacy benefits program to receive (or not) several mailings on osteoporosis and fracture prevention.

Such a factorial design allows for evaluation of the separate effects of the 2 interventions. A limitation of such factorial designs is that the interventions are generally assumed to have independent, additive effects. Such trials generally have limited power to detect interactions that occur, for example, if both interventions are necessary for patients to benefit.[11] This factorial design is more appropriate if the multiple randomized treatments are directed toward different outcomes (eg, improved prescribing by doctors and greater awareness of treatment options, benefits, and risks by patients).

THE DESIGN EFFECT

The design of any multipractice study must consider the degree of clustering among patients within a practice. The intracluster correlation coefficient (also called the intraclass correlation coefficient and often denoted by the Greek letter ρ) quantifies the amount of agreement in a characteristic between 2 people in the same practice.[12] Many common statistical tests make the assumption that units of analysis are independent and identically distributed. This assumption is violated when patients within a practice are more alike than those from other practices. Usually, assuming the same total number of patients in a given study, the variance of the sample mean of a characteristic measured in a study with more than 1 patient sampled from each of several practices is greater than the variance of the sample mean of that characteristic in a study with 1 patient included per practice.

Specifically, if we observe an outcome Y in each of m patients within each of k practices, the intracluster correlation coefficient quantifies the strength of agreement in Y between 2 patients treated in the same practice. It is the ratio of the variance between practices, denoted σ_B^2, divided by the sum of the variance between practices plus the variance among patients within a practice, denoted σ_W^2 (ie, $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$). To estimate ρ, we calculate the between mean square (BMS) as $BMS = \sum_{i=1}^{k} m(\bar{y}_i - \bar{y})^2/(k - 1)$, where $\bar{y}_i$ is the mean of Y in the ith practice, and $\bar{y}$ is the overall mean of Y across all practices; and the within mean square (WMS) as $\mathrm{WMS} = \sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ij} - \bar{y}_i)^2/(mk - k)$ where y_{ij} is the value of Y in the jth patient from the ith practice. Then, the estimated intracluster correlation is:

$$\hat{\rho} = (\mathrm{BMS} - \mathrm{WMS})/(\mathrm{BMS} + (m - 1)\mathrm{WMS}).$$

If the number of patients in each practice varies, so that there are m_i patients in the ith practice, then m is replaced by m_i in the formula for BMS, mk is replaced by the total number of patients across all practices in the formula for WMS, and m is replaced by m_0 in the formula for $\hat{\rho}$, where $m_0 = (\sum_{i=1}^{k} m_i - \sum_{i=1}^{k} m_i^2/\sum_{i=1}^{k} m_i)/(k - 1)$.

Often the goal of a study is to estimate the average value of Y among all subjects in a given treatment group. If we let $\sigma^2 = \sigma_B^2 + \sigma_W^2$ denote the variance of a single observation of Y, then the variance of the mean is $var(\bar{Y}) = [1 + (m - 1)\rho]\sigma^2/(km)$.

The term $1 + (m - 1)\ \rho$ is called the design effect because it describes the increase in the variance of the mean for a study with m patients in each of k practices compared with the variance of the mean for a study that included the same total number of patients but only 1 patient per practice. If Y is dichotomous and the prevalence of Y follows a binomial distribution with mean p, then the above formula holds with $p\ (1 - p)$ in place of σ^2, so the design effect also describes the increase in variance associated with clustering with discrete outcomes.

In the common situation with a variable number of subjects per cluster, the average $\bar{m}$ is substituted for m, although this will slightly underestimate the design effect.[13] A positive value of ρ, and hence a positive design effect, indicates the need to design a cluster-randomized trial with a higher sample size than an individually randomized trial would require. A review of previous cluster-randomized trials found that investigators often ignored the design effect in their analyses, which led to erroneous P values, but they more often ignored the design effect in sample size calculations, so that power was inadequate to detect the hypothesized effect.[14]

Planning for a cluster-randomized trial requires an a priori estimate of the design effect. In many settings, one can use standard approaches for estimating sample size, assuming no clustering, and then increase the number of subjects by a factor equal to the design effect.[13]

As an example, consider a trial to evaluate an intervention to improve screening and care for osteoporosis within a large insurance plan in New Jersey.[15] The intervention was focused on women 65 years of age or older, or any enrollee 45 years or older with a prior fracture or recent use of glucocorticoids. The intervention included education and reminders to primary care physicians and mailings and calls to their at-risk patients.

The observed design effect in this trial was estimated from 1973 patients in 435 primary care practices, based on formulas presented above. Table 1 shows estimated intracluster correlation coefficients and design effects for the 3 primary endpoints observed during 10-month follow-up: receipt

TABLE 1. Observed Intracluster Correlations and Design Effects for Outcomes in a Trial of Strategies to Improve Osteoporosis Care: 1973 Patients in 435 Primary Care Practices[15]

Outcome	Intracluster Correlation	Design Effect
Bone mineral density test	0.042	1.15
Prescription for osteoporosis	0.025	1.09
Either test or prescription	0.035	1.12

TABLE 2. Impact of Cluster Size *m* and Intracluster Correlation (ICC) on the Design Effect

	ICC				
m	0.01	0.02	0.03	0.04	0.05
4	1.03	1.06	1.09	1.12	1.15
10	1.09	1.18	1.27	1.36	1.45
20	1.19	1.38	1.57	1.76	1.85
50	1.49	1.98	2.47	2.96	3.45
200	2.99	4.98	6.97	8.96	10.95

of a bone mineral density test; a filled prescription for an osteoporosis medication; or a composite consisting of either of these 2 outcomes. The range of observed within-practice correlations, from 0.025 to 0.042, is consistent with typical intracluster correlations observed in other studies within practices.[16] With the observed average of 4.5 patients per practice in this trial of strategies to improve osteoporosis care,[15] the calculated design effects indicate that sample size must be increased between 9% and 15%, compared with a study that recruited 1 patient per practice.

As illustrated in Table 2 for the range of intracluster correlations observed in this study,[15] the observed design effect will vary according to the average number of patients included per cluster. Increasing the number of patients in each cluster leads to a larger design effect for a fixed intracluster correlation, but this may still be a cost-effective strategy if the marginal cost per added patient in a cluster is low. In this example, the major costs involve contact and education of physicians, with little additional cost for inclusion of extra patients within a practice.

Adams and colleagues[16] have argued that only limited information has been published on intracluster correlations and design effects in multipractice studies. Because appropriate estimates are often unavailable at the time a cluster-randomized trial is planned, investigators often have to estimate study power for a range of possible design effects and guess at the most reasonable values. However, when cluster-randomized trials are conducted within health care systems with available automated information on patient characteristics and use of services, researchers can use this information to obtain reliable estimates of the design effect as they plan a study. Although the example we presented considered the observed outcome during a cluster-randomized trial, a pretrial estimate could also be obtained from historical information in archived claims data on the same target population. Such historical information is generally available when interventions are coordinated within systems with regular data monitoring. Better knowledge of the anticipated intracluster correlation in the outcome at the time of study design will increase the likelihood that the trial is adequately powered to answer the question of interest.

FOCUS ON THOSE WHO WILL ADHERE AND BENEFIT

Randomized trials use eligibility criteria to identify those people who will most likely benefit from an intervention and to exclude those with contraindications. Often trials have extensive eligibility and exclusion criteria, although simple, wide entry criteria can increase enrollment and enhance generalizability.[17] Narrow criteria can limit understanding of the applicability of interventions in groups of patients that might benefit substantially. For example, older people and women were frequently excluded from past clinical trials of treatments for myocardial infarction, leaving providers to wonder about the value of treatments in the age group with the highest event rates.[18]

Wide eligibility criteria are particularly relevant for evaluation of quality improvement interventions because of the need to demonstrate broad applicability of findings. Nonetheless, some eligibility criteria are usually necessary to focus on those who will benefit from the intervention. For example, interventions to improve preventive care may have limited value for patients at the end of life, such as those with metastatic cancer for whom palliative care is more relevant. Similarly, an intervention involving direct patient contact may have little impact among demented patients. Inclusion in the study population of people with little likelihood of compliance with study procedures will decrease the observed effect and dilute the power of the study. Once such people are randomized, they must be included in intention to treat analyses. For this reason, individually randomized trials often include a run-in period to evaluate compliance before randomization. Use of a run-in period can reduce generalizability but enhance the internal validity of the study.[19] In the Physicians' Health Study, compliance during the run-in period was a powerful determinant of compliance during the trial.[20] Because increased compliance enhances internal validity and internal validity is a prerequisite for generalizability, eligibility criteria that focus on people most likely to benefit from interventions often have a net positive effect on the value of a trial.

Implementation of eligibility criteria is challenging when practices are the unit of randomization and individual patients are not personally evaluated before enrollment. However, when a trial is embedded in a health care system with automated claims data, substantial information may be available to specify useful eligibility criteria in the design of the trial. For example, the intervention to improve care for osteoporosis described above focused on at-risk patients including women age 65 years or older and those with a history of fracture or recent use of glucocorticoids. Other eligibility (or inclusion) criteria based on claims data could be added to identify patients most likely to respond to interventions such as this.

UNBALANCED COVARIATES AND BALANCING SCORES

With reasonable sample sizes, randomization will tend to yield balance between groups in both observed and unobserved covariates measured at the level of randomization. However, with cluster randomization at the level of practices or providers, characteristics of patients might still differ between groups if a few providers have an unusual mix of patients.

The trial of interventions to improve care for osteoporosis illustrates this phenomenon.[15] Table 3 shows that char-

TABLE 3. Baseline Characteristics of Physicians in the Primary Analytic Population: Trial of Interventions to Improve Osteoporosis Care[15]

Physician Characteristics	Intervention	Control	P*
No.	223	212	N/A
At-risk patients, median (IQR)	4 (2–5)	3 (2–5)	0.08
Age, yr, median (IQR)	49 (44–55)	49 (44–56)	0.7
Female, N (%)	40 (18)	33 (16)	0.5
Family medicine, N (%)	87 (39)	69 (33)	0.3
Internal medicine, N (%)	99 (44)	101 (48)	—
Subspecialty, N (%)	37 (17)	42 (20)	—

*P value based on rank-sum test for ordered characteristics, and χ^2 test for gender comparison and comparison of the distribution of medical specialties between groups.

IQR indicates interquartile range; N/A, not applicable.

TABLE 4. Baseline Characteristics of Patients in the Primary Analytic Population: Trial of Interventions to Improve Osteoporosis Care[15]

Patient Characteristics	Intervention	Control	P*
No.	997	976	—
Age, yr, median (IQR)	68 (65–72)	69 (66–74)	0.004
Female, N (%)	895 (90)	922 (94)	<0.001
Medications, N median (IQR)	12 (7–18)	11 (7–18)	0.5
Physician visits, N median (IQR)	13 (7–22)	13 (6–22)	0.8
Fractures, N (%)	134 (13)	95 (10)	0.02

*P-values for comparison of dichotomous characteristics from a logistic regression model accounting for clustering within practices by generalized estimating equations; for ordinal and continuous characteristics, a normal, random-effects model accounted for clustering within practices.

acteristics measured at the level of the provider, the unit of randomization in this trial, were generally balanced. By contrast, Table 4 shows that some characteristics measured at the level of the patient were not balanced, with a greater percentage of women, a slightly older average age in the control group, and a greater percentage of patients with a history of fracture in the intervention group. This example agrees with the experience of Puffer and colleagues[2] who reviewed 36 published cluster-randomized trials; they found little evidence for imbalance at the cluster level but potential risk of bias at the individual level in 39% of trials.

Thus, imbalance between groups in individual-level covariates poses a substantial threat to the validity of cluster-randomized trials. To some extent, such imbalances can be addressed at the time of analysis. However, addressing potential imbalances at the time of study design will simplify the presentation of results, allow randomization inference, and possibly provide better control for confounding. Raab and Butcher[21] have discussed strategies for balancing such covariates by design. With information on covariates predictive of outcome collected upon recruitment of clusters, investigators can apply several approaches to balance these covariates across treatment groups.

If one seeks to balance a few categorical covariates, then separate randomizations of clusters can be performed within blocks defined by the cross-classification of these covariates. Alternatively, one can improve balance through sequential treatment assignment where at the time of each randomization the distribution of covariates between treatment groups is compared and treatment assignment is dynamically chosen to minimize imbalances between covariate patterns.[22] Thus, for a 2-arm trial, the probability that a cluster will be assigned to a given arm will generally not be 0.5, as it will be modified to favor an allocation that would lessen differences in prespecified covariates among all clusters randomized to date. This latter approach, also called the minimization technique, is particularly useful if the number of covariate patterns approaches the number of units to be randomized.[23] Steptoe and colleagues[24] used minimization to balance several covariates in a cluster-randomized trial of behavioral counseling conducted in 20 general practices. Another approach is to consider only designs with small average differences between treatment groups in some set of covariates. This approach was taken by Henderson and colleagues[25] in a cluster-randomized trial of a sex education program in 25 schools in Scotland.

An alternative approach is possible when trials are conducted within health care systems with claims data. Investigators can use historical information from these data to identify characteristics predictive of the outcome of interest. If many characteristics are predictive, a multivariate risk score can be developed, and this score can account for clustering within practices. Then, in the design of a cluster-randomized trial, practices in each treatment group can be balanced on their distributions of these overall predictive scores. Use of such a score is akin to balancing on a propensity score or a multivariate risk score.[26]

To illustrate the approach, consider the possible use of a balancing score in the context of the trial of interventions to improve care for osteoporosis.[15] The administrative claims data for the potentially eligible patients in this trial contained information on use of osteoporosis medications and screening tests for osteoporosis in the period before randomization. Based on these data, we developed a predictive model for such baseline use including the variables shown in Table 4. Then, for each eligible physician, we obtained the average predicted probability of osteoporosis care for all eligible patients, based on their observed covariates. We next divided physicians into 4 groups according to their average predicted probability of osteoporosis care, and also formed 3 strata according to the number of eligible patients. We then randomly assigned physicians to intervention or control group in blocks of 4 within the 12 strata formed by the cross-classification of these measures of average treatment probability and practice size. With this alternative randomization, baseline covariates of patients displayed better balance between treatment groups (Table 5), suggesting the utility of this approach.

SUMMARY

Availability of automated information on providers and patients in a health care system can provide critical information to improve the design of a cluster-randomized trial to evaluate a quality improvement intervention within that sys-

TABLE 5. Comparison of Baseline Characteristics of Patients in the Primary Analytic Population After Alternative Randomization Blocked on a Predictive Score and Practice Size: Trial of Interventions to Improve Osteoporosis Care[15]

Patient Characteristics	Intervention	Control	*P**
No.	1010	963	—
Age, yr, median (IQR)	69 (66–73)	68 (65–73)	0.5
Female, N (%)	936 (93)	881 (91)	0.4
Medications, N median (IQR)	12 (7–18)	11 (7–18)	0.6
Physician visits, N median (IQR)	12 (6–21)	14 (7–23)	0.07
Fractures, N (%)	123 (12)	106 (11)	0.5

**P*-values for comparison of dichotomous characteristics from a logistic regression model accounting for clustering within practices by generalized estimating equations; for ordinal and continuous characteristics, a normal, random-effects model accounted for clustering within practices.

tem. A first decision in study design is the choice of the unit of randomization, and data on sharing of patients and similarity of care among providers within a practice can help investigators decide whether to randomize individual providers or practices. A challenge in this regard is the need to define and validate algorithms to assign patients to a unique provider when they see several providers within a practice.

The necessary sample size to provide a definitive answer to a question of interest depends on the intracluster correlation among patients within a practice. Preintervention estimates of the range of design effects that might be expected can yield realistic sample size determinations. Some patients have a decreased likelihood of benefit and may also be less likely to respond to an intervention. A priori restriction to those with a reasonable likelihood of adherence and benefit can enhance the validity of an intervention trial. Imbalances in subject-specific covariates are especially likely in cluster-randomized trials. However, investigators can minimize such imbalances with a design that uses balancing scores or other strategies based on available automated data.

Certainly there are limitations as well as strengths associated with the use of automated claims data to design cluster-randomized trials, and also to assess outcomes within enrolled populations.[27] Some potentially important covariates that influence outcomes and might be targets for balancing by design may not be available in claims data sets. For example, patient characteristics such as weight, cigarette smoking, alcohol consumption, and physical activity are often not available in claims data. Many claims data sets do not contain actual laboratory values. However, automated data sets may be optimal for assessment of history of system use, prior diagnoses, and use of medications. Outcomes identified through claims data are subject to measurement error,[28] and privacy concerns may preclude investigators from receipt of detailed medical records needed to confirm diagnoses. However, for health services outcomes such as receipt of services or filled prescriptions, automated databases may be optimal because reimbursement to providers requires a filed claim. For such outcomes, carefully constructed claims data sets may serve as the gold standard for outcome assessment.

REFERENCES

1. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ*. 2004;328:702–708.
2. Puffer S, Torgerson DJ, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ*. 2003;327:785–789.
3. Hutton JL. Are distinctive ethical principles required for cluster randomized trials? *Stat Med*. 2001;20:473–488.
4. Zelen M. Compliance, bias and power in clinical trials: reply. *Biometrics*. 1991;47:778–779.
5. Rosner B. *Fundamentals of Biostatistics*. 6th ed. Belmont, CA: Thomson; 2006:423.
6. Solomon DH, van Houten L, Glynn RJ, et al. Academic detailing to improve use of broad-spectrum antibiotics at an academic medical center. *Arch Intern Med*. 2001;161:1697–1902.
7. Avorn J, Soumerai S. Improving drug-therapy decisions through educational outreach: a randomized controlled trial of academically based detailing. *N Engl J Med*. 1983;308:1457–1463.
8. Soumerai S, Avorn J. Principles of educational outreach (academic detailing) to improve clinical decision-making. *JAMA*. 1990;263:549–556.
9. Brookhart MA, Solomon DH, Wang P, et al. Explained variation in a model of therapeutic decision making is partitioned across patient, physician, and clinic factors. *J Clin Epidemiol*. 2006;59:18–25.
10. Solomon DH, Brookhart MA, Polinski J, et al. Osteoporosis action: design of the Healthy Bones Project Trial. *Contemp Clin Trials*. 2005;26:78–94.
11. Friedman LM, Furburg CD, DeMets DL. *Fundamentals of Clinical Trials*. 3rd ed. St Louis, MO: Mosby Year Book; 1996.
12. Fleiss JL. *The Design and Analysis of Clinical Experiments*. New York, NY: John Wiley & Sons; 1986.
13. Donner A, Klar N. *Design and Analysis of Cluster Randomized Trials*. London, England: Arnold; 2000.
14. Simpson JM, Klar N, Donner A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *Am J Public Health*. 1995;85:1378–1382.
15. Solomon DH, Polinski JM, Stedman M, et al. Improving care of patients at-risk for osteoporosis: a randomized controlled trial. *J Gen Intern Med*. 2007;22:362–367.
16. Adams G, Gulliford MC, Uloumunne OC, et al. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol*. 2004;57:785–794.
17. Yusuf S, Held P, Teo KK, et al. Selection of patients for randomized controlled trials: implications of wide or narrow eligibility criteria. *Stat Med*. 1990;9:73–83.
18. Gurwitz JH, Col NF, Avorn J. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *JAMA*. 1992;268:1417–1422.
19. Glynn RJ, Buring JE, Hennekens CH. Concerns about run-in periods in randomized trials (letter). *JAMA*. 1998;279:1526–1527.
20. Glynn RJ, Buring JE, Manson JE, et al. Adherence to aspirin in the prevention of myocardial infarction: the Physicians' Health Study. *Arch Intern Med*. 1994;154:2649–2657.
21. Raab GM, Butcher I. Balance in cluster randomized trials. *Stat Med*. 2001;20:351–365.
22. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*. 1975;31:103–115.
23. MacRae KD. Minimisation: the platinum standard for trials? *BMJ*. 1998;317:362–363.
24. Steptoe A, Doherty S, Rink E, et al. Behavioural counseling in general practice for the promotion of healthy behaviour among adults at increased risk of coronary heart disease: randomised trial. *BMJ*. 1999;319:943–948.
25. Henderson M, Wight D, Raab GM, et al. Impact of a theoretically based sex education programme (SHARE) delivered by teachers on NHS registered conceptions and terminations: final results of a cluster randomised trial. *BMJ*. 2007;334:133.
26. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*. 2006;98:253–259.
27. Schneeweiss S, Avorn J. A review of use of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58:323–337.
28. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. *J Clin Epidemiol*. 2004;57:131–141.