

Uses and abuses of screening tests

David A Grimes, Kenneth F Schulz

Screening tests are ubiquitous in contemporary practice, yet the principles of screening are widely misunderstood. Screening is the testing of apparently well people to find those at increased risk of having a disease or disorder. Although an earlier diagnosis generally has intuitive appeal, earlier might not always be better, or worth the cost. Four terms describe the validity of a screening test: sensitivity, specificity, and predictive value of positive and negative results. For tests with continuous variables—eg, blood glucose—sensitivity and specificity are inversely related; where the cutoff for abnormal is placed should indicate the clinical effect of wrong results. The prevalence of disease in a population affects screening test performance: in low-prevalence settings, even very good tests have poor predictive value positives. Hence, knowledge of the approximate prevalence of disease is a prerequisite to interpreting screening test results. Tests are often done in sequence, as is true for syphilis and HIV-1 infection. Lead-time and length biases distort the apparent value of screening programmes; randomised controlled trials are the only way to avoid these biases. Screening can improve health; strong indirect evidence links cervical cytology programmes to declines in cervical cancer mortality. However, inappropriate application or interpretation of screening tests can rob people of their perceived health, initiate harmful diagnostic testing, and squander health-care resources.

Screening is a double-edged sword, sometimes wielded clumsily by the well-intended. Although ubiquitous in contemporary medical practice, screening remains widely misunderstood and misused. Screening is defined as tests done among apparently well people to identify those at an increased risk of a disease or disorder. Those identified are sometimes then offered a subsequent diagnostic test or procedure, or, in some instances, a treatment or preventive medication.¹ Looking for additional illnesses in those with medical problems is termed case finding;^{2,3} screening is limited to those apparently well.

Screening can improve health. For example, strong indirect evidence lends support to cytology screening for cervical cancer. Insufficient use of this screening method accounts for a large proportion of invasive cervical cancers in industrialised nations.⁴ Other beneficial examples include screening for hypertension in adults; screening for hepatitis B virus antigen, HIV-1, and syphilis in pregnant women; routine urine culture in pregnant women at 12–16 weeks' gestation; and measurement of phenylalanine in newborns.⁵ However, inappropriate screening harms healthy individuals and squanders precious resources. The nearly universal antenatal screening for gestational diabetes (a diagnosis in search of a disease)⁶ in the USA⁷ exemplifies the widespread confusion about the nature and aim of screening. Here, we review the purposes of screening, the selection of tests, measurement of validity, the effect of prevalence on test outcome, and several biases that can distort interpretation of tests.

Ethical implications

What are the potential harms of screening?

Screening differs from the traditional clinical use of tests in several important ways. Ordinarily, patients consult with clinicians about complaints or problems; this prompts testing to confirm or exclude a diagnosis.⁸

Lancet 2002; **359**: 881–84

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (D A Grimes MD, K F Schulz PhD)

Correspondence to: Dr David A Grimes
(e-mail: dgrimes@fhi.org)

Because the patient is in pain and requests our help, the risk and expense of tests are usually deemed acceptable by the patient. By contrast, screening engages apparently healthy individuals who are not seeking medical help (and who might prefer to be left alone). Alternatively, consumer-generated demand for screening, such as for osteoporosis and ovarian cancer, might lead to expensive programmes of no clear value.⁹ Hence, the cost, injury, and stigmatisation related to screening are especially important (though often ignored in our zeal for earlier diagnosis); the medical and ethical standards of screening should be, correspondingly, higher than with diagnostic tests.¹⁰ Bluntly put: every adverse outcome of screening is iatrogenic and entirely preventable.

Screening has a darker side that is often overlooked.² It can be inconvenient (the O'Sullivan screen for gestational diabetes), unpleasant (sigmoidoscopy or colonoscopy), and expensive (mammography). For example, a recent Markov model revealed that new screening tests for cervical cancer that are more sensitive than the Papanicolaou test (and thus touted as being better) will drive up the average cost of detecting an individual with cancer.¹¹ Paradoxically, these higher costs could make screening unattainable by poor women who are at highest risk.⁴ The net effect might be more instances of cancer.

A second wave of injury can arise after the initial screening insult: false-positive results and true-positive results leading to dangerous interventions.² Although the stigma associated with correct labeling of people as ill might be acceptable, those incorrectly labeled as sick suffer as well. For example, labeling productive steelworkers as being hypertensive led to increased absenteeism and adoption of a sick role, independent of treatment.^{12,13} More recently, women labeled as having gestational diabetes reported deterioration in their health and that of their infants over the 5 years after diagnosis.¹⁴ By what right do clinicians rob people of their perceived health, and for what gain?²

Screening can also lead to harmful treatment. Treatment of hyperlipidaemia with clofibrate several decades ago provides a sobering example. Treatment of the cholesterol count (a risk factor, rather than an illness itself) inadvertently led to a 17% increase in mortality among middle-aged men given the drug.² This screening

misadventure cost the lives of more than 5000 men in the USA alone.² Because of these mishaps, reviews of screening practices have recommended that clinicians be more selective.^{5,15}

Criteria for screening

If a test is available, should it be used?

The availability of a screening test does not imply that it should be used. Indeed, before screening is done, the strategy must meet several stringent criteria. One checklist separates criteria in three parts: the disease, the policy, and the test.¹ The disease should be medically important and clearly defined, and its prevalence reasonably well known. The natural history should be known, and an effective intervention must exist. Concerning policy, the screening programme must be cost effective, facilities for diagnosis and treatment must be readily available, and the course of action after a positive result must be generally agreed on and acceptable to those screened. Finally, the test must do its job. It should be safe, have a reasonable cut-off level defined, and be both valid and reliable. The latter two terms, often used interchangeably, are distinct. Validity is the ability of a test to measure what it sets out to measure, usually differentiating between those with and without the disease. By contrast, reliability indicates repeatability. For example, a bathroom scale that consistently measures 2 kg heavier than a hospital scale (the gold standard) provides an invalid but highly reliable result.

Although an early diagnosis generally has intuitive appeal, earlier might not always be better. For example, what benefit would accrue (and at what cost) from early diagnosis of Alzheimer's disease, which to date has no effective treatment? Sackett and colleagues² have proposed a pragmatic checklist to help decide when (or if) seeking a diagnosis earlier than usual is worth the expense and bother. Does early diagnosis really benefit those screened, for example, in survival or quality of life? Can the clinician manage the additional time required to confirm the diagnosis and deal with those diagnosed before symptoms developed? Will those diagnosed earlier comply with the proposed treatment? Has the effectiveness of the screening strategy been established objectively?^{5,15} Finally, are the cost, accuracy, and acceptability of the test clinically acceptable?

Assessment of test effectiveness

Is the test valid?

For over half a century,¹⁶ four indices of test validity have been widely used: sensitivity, specificity, and predictive values of positive and negative. Although clinically useful (and far improved over clinical hunches), these terms are predicated on an assumption that is often clinically unrealistic—ie, that all people can be dichotomised as ill or well. (Indeed, one definition of an epidemiologist is a person who sees the entire world in a 2×2 table.) Often, those tested simply do not fit neatly into these designations: they might be possibly ill, early ill, probably well, or some other variant. Likelihood ratios, which incorporate varying (not just dichotomous) degrees of test results, can be used to refine clinicians' judgments about the probability of disease in a particular person.

For simplicity, however, assume a population has been tested and assigned to the four mutually exclusive cells in figure 1. Sensitivity, sometimes termed the detection rate,¹⁰ is the ability of a test to find those with the disease. All those with disease are in the left column. Hence, the sensitivity is simply those correctly identified by the test (a) divided by all those sick (a+c). Specificity denotes the

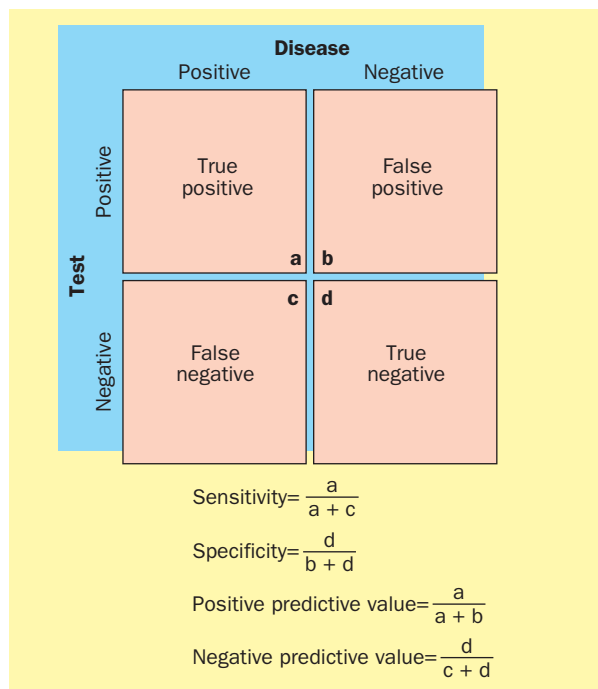


Figure 1: Template for calculation of test validity

ability of a test to identify those without the condition. Calculation of this proportion is trickier, however. By analogy to sensitivity, many assume (incorrectly) that the formula here is $b/(b+d)$. However, the numerator for specificity is cell d (the true negatives), which is divided by all those healthy (b+d).

Although sensitivity and specificity are of interest to public-health policymakers, they are of little use to the clinician. Stated alternatively, sensitivity and specificity (population measures) look backward (at results gathered over time).⁸ Clinicians have to interpret test results to those tested. Thus, what clinicians need to know are the predictive values of the test (individual measures, which look forward). To consider predictive values, one needs to shift the orientation in figure 1 by 90 degrees: predictive values work horizontally (rows), not vertically (columns). In the top row are all those with a positive test, but only those in cell a are sick. Thus, the predictive value positive is $a/(a+b)$. The “odds of being affected given a positive result (OAPR)” is the ratio of true positives to false positives, or a to b.¹⁰ For example, in figure 1, the OAPR is 75/5, or 17/1. This corresponds to a positive predictive value of 89%. Advocates of use of the OAPR note that these odds better describe test effectiveness than do probabilities (predictive values). In the bottom row of figure 1 are those with negative tests, but only those in cell d are free of disease. Hence, the predictive value negative is $d/(c+d)$.

Learning (and promptly forgetting) these formulas was an annual ritual for many of us in our clinical training. If readers understand the definitions above and can recall the 2×2 table shell, then they can quickly figure out these formulas when needed. As a mnemonic, disease goes at the top of the table shell, since it is our top priority. By default, test goes on the left border.

Through the years, researchers have tried to simplify these four indices of test validity by condensing them into a single term.⁸ However, none adequately depicts the important trade-offs between sensitivity and specificity that generally arise. An example is diagnostic accuracy, which is the proportion of correct results.³ It is the sum of

the correctly identified ill and well divided by all those tested, or $(a+d)/(a+b+c+d)$. Cells b and c are noise in the system. Another early attempt, Youden's J, is simply the predictive value positive plus the predictive value negative minus one.¹⁷ The range of values extends from zero (for a coin toss with no predictive value) to 1.0, where predictive values of both positive and negative tests are perfect.

Trade-offs between sensitivity and specificity

Where should the cut-off for abnormal be?

The ideal test would perfectly discriminate between those with and without the disorder. The distributions of test results for the two groups would not overlap. More commonly in human biology, test values for those with and without a disease overlap, sometimes widely.¹⁸ Where one puts the cut-off defining normal versus abnormal determines the sensitivity and specificity. For any continuous outcome measurement—for example, blood pressure, intraocular pressure, or blood glucose—the sensitivity and specificity of a test will be inversely related. Figure 2 shows that placing the cut-off for abnormal blood glucose at point X produces perfect sensitivity; this low cut-off identifies all those with diabetes. However, the trade-off is poor specificity: those in the part of the healthy distribution in pink and purple are incorrectly identified as having abnormal values. Placing the cut-off higher at point Z yields the opposite result: all those healthy are correctly identified (perfect specificity), but the cost here is missing a proportion of ill individuals (portion of the diabetic distribution in purple and blue). Placing the cut-off at point Y is a compromise, mislabeling some healthy people and some people with diabetes.

Where the cut-off should be depends on the implications of the test, and receiver-operator characteristic curves are useful in making this decision.¹⁹ For example, screening for phenylketonuria in newborns places a premium on sensitivity rather than on specificity; the cost of missing a case is high, and effective treatment exists. The downside is a large number of false-positive tests, which cause anguish and further testing. By contrast, screening for breast cancer should favour specificity over sensitivity, since further assessment of those tested positive entails costly and invasive biopsies.²⁰

Prevalence and predictive values

Can test results be trusted?

A badly understood feature of screening is the potent effect of disease prevalence on predictive values.

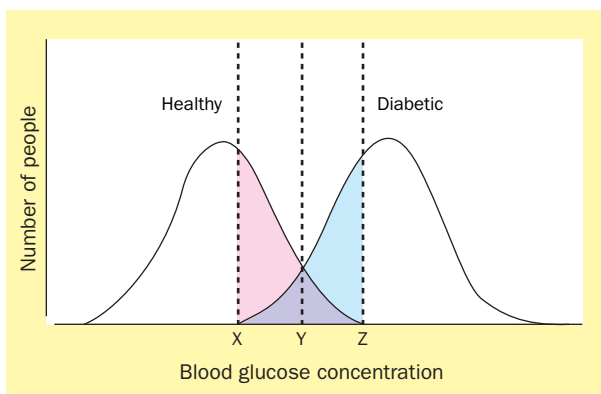


Figure 2: Hypothetical distribution of blood glucose concentrations in people with and without diabetes

Setting cut-off for abnormal at X yields perfect sensitivity at the expense of specificity. Setting cut-off at Z results in perfect specificity at the cost of lower sensitivity. Cut-off Y is a compromise.

Clinicians must know the approximate prevalence of the condition of interest in the population being tested; if not, reasonable interpretation is impossible. Consider a new PCR test for chlamydia, with a sensitivity of 0.98 and specificity of 0.97 (a superb test). As shown in the left panel of figure 3, a doctor uses the test in a municipal sexually transmitted disease clinic, where the prevalence of *Chlamydia trachomatis* is 30%. In this high-prevalence setting, the predictive value of a positive test is high, 93%—ie, 93% of those with a positive test actually have the infection.

Impressed with the new test, the doctor now takes it to her private practice in the suburbs, which has a clientele that is mostly older than age 35 years (figure 3, right panel). Here, the prevalence of chlamydial infection is only 3%. Now the same excellent test has a predictive positive value of only 0.50. When the results of the test are positive, what should the doctor tell the patient, and what, in turn, should the patient tell her husband? Here, flipping a coin has the same predictive positive value (and is considerably cheaper and simpler than searching for bits of DNA). This message is important, yet not widely understood: when used in low-prevalence settings, even excellent tests have poor predictive positive value. The reverse is true for negative predictive values, which are nearly perfect in figure 3. Although failing to diagnose sexually transmitted diseases can have important health implications, incorrectly labeling people as infected can wreck marriages and damage lives.

Tests in combination

Should a follow-up test be done?

Clinicians rarely use tests in isolation. Few tests have high sensitivity and specificity, so a common approach is to do tests in sequence. In the instance of syphilis, a sensitive (but not specific) reagin test is the initial screen. Those who test positive then get a second, more specific test, a diagnostic treponemal test. Only those who test positive on both receive the diagnosis. This strategy generally increases the specificity compared with a single test and limits the use of the more expensive treponemal test.²⁰ Testing for HIV-1 is an analogous two-step procedure.

Alternatively, tests can be done in tandem (parallel or simultaneous testing).^{3,21} For example, two different tests might both have poor sensitivity, but one might be better at picking up early disease, whereas the other is better at

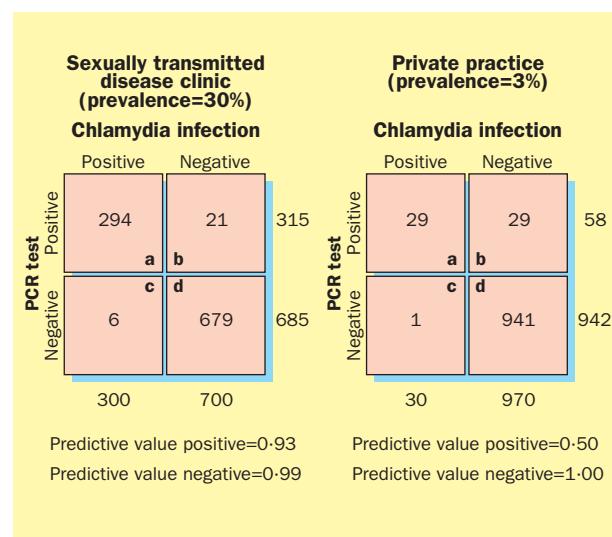


Figure 3: Predictive values of a PCR test for *Chlamydia trachomatis* in high-prevalence and low-prevalence settings

identifying late disease. A positive result from either test would then lead to diagnostic assessment. This approach results in higher sensitivity than would arise with either test used alone.

Benefit or bias?

Does a screening programme really improve health?

Even worthless screening tests seem to have benefit.² This cruel irony underlies many inappropriate screening programmes used today. Two common pitfalls lead to the conclusion that screening improves health; one is an artifact and the other a reflection of biology.

Lead-time bias

Lead-time bias refers to a spurious increase in longevity associated with screening. For example, assume that mammography screening leads to cancer detection 2 years earlier than would have ordinarily occurred, yet the screening does not prolong life. On average, women with breast cancer detected through screening live 2 years longer than those with cancers diagnosed through traditional means. This gain in longevity is apparent and not real: this hypothetical screening allows women to live 2 years longer with the knowledge that they have cancer, but does not prolong survival, an example of zero-time shift.²

Length bias

Length bias is more subtle than lead-time bias: the longevity association is real, but indirect. Assume that community-based mammography screening is done at 10-year intervals. Women whose breast cancers were detected through screening live 5 years longer on average from cancer initiation to death than those whose cancers were detected through usual means. That screening is associated with longer survival implies clear benefit. However, in this hypothetical example, this benefit indicates the inherent variability in cancer growth rates and not a benefit of screening. Women with indolent, slow-growing cancers are more likely to live long enough to be identified in decennial screening. Conversely, those with rapidly progressing tumours are less likely to survive until screening.

The only way to avoid these pervasive biases is to do randomised controlled trials and then to assess age-specific mortality rates for those screened versus those not screened.¹⁰ Moreover, the trials must be done well. The quality of published trials of mammography screening has raised serious questions about the utility of this massive and hugely expensive enterprise.²²⁻²⁴

Conclusion

Screening can promote or impair health, depending on its application. Unlike a diagnostic test, a screening test is done in apparently healthy people, which raises unique ethical concerns. Sensitivity and specificity tend to be inversely related, and choice of the cut-off point for abnormal should indicate the implications of incorrect results. Even very good tests have poor predictive value positive when applied to low-prevalence populations.

Lead-time and length bias exaggerate the apparent benefit of screening programmes, underscoring the need for rigorous assessment in randomised controlled trials before use of screening programmes.

Acknowledgments

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of this material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

- Cuckle HS, Wald NJ. Principles of screening. In: Antenatal and neonatal screening. Oxford: Oxford University Press, 1984: 1-22.
- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine, 2nd edn. Boston: Little, Brown and Company, 1991.
- Lang TA, Secic M. How to report statistics in medicine. Philadelphia: American College of Physicians, 1997.
- Sawaya GF, Grimes DA. New technologies in cervical cytology screening: a word of caution. *Obstet Gynecol* 1999; **94**: 307-10.
- US Preventive Services Task Force. Guide to clinical preventive services, 2nd edn. Baltimore: Williams and Wilkins, 1996.
- Enkin M, Keirse MJNC, Neilson J, et al (eds). A guide to effective care in pregnancy and childbirth, 3rd edn. Oxford: Oxford University Press, 2000.
- Gabbe S, Hill L, Schmidt L, Schulkin J. Management of diabetes by obstetrician-gynecologists. *Obstet Gynecol* 1998; **91**: 643-47.
- Feinstein AR. Clinical biostatistics XXXI: on the sensitivity, specificity, and discrimination of diagnostic tests. *Clin Pharmacol Ther* 1975; **17**: 104-16.
- NIH Consensus Development Panel on Ovarian Cancer. Ovarian cancer: screening, treatment, and follow-up. *JAMA* 1995; **273**: 491-97.
- Wald N, Cuckle H. Reporting the assessment of screening and diagnostic tests. *Br J Obstet Gynaecol* 1989; **96**: 389-96.
- Myers ER, McCrory DC, Subramanian S, et al. Setting the target for a better cervical screening test: characteristics of a cost-effective test for cervical neoplasia screening. *Obstet Gynecol* 2000; **96**: 645-52.
- Haynes RB, Sackett DL, Taylor DW, Gibson ES, Johnson AL. Increased absenteeism from work after detection and labeling of hypertensive patients. *N Engl J Med* 1978; **299**: 741-44.
- Taylor DW, Haynes RB, Sackett DL, Gibson ES. Longterm follow-up of absenteeism among working men following the detection and treatment of their hypertension. *Clin Invest Med* 1981; **4**: 173-77.
- Feig DS, Chen E, Naylor CD. Self-perceived health status of women three to five years after the diagnosis of gestational diabetes: a survey of cases and matched controls. *Am J Obstet Gynecol* 1998; **178**: 386-93.
- Canadian Task Force on the Periodic Health Examination. The Canadian guide to clinical preventive care. Ottawa: Minister of Supply and Services Canada, 1994.
- Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Pub Health Rep* 1947; **62**: 1432-49.
- Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3**: 32-35.
- Griffith CS, Grimes DA. The validity of the postcoital test. *Am J Obstet Gynecol* 1990; **162**: 615-20.
- Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987; **6**: 411-23.
- Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Company, 1987.
- Riegelman RK, Hirsch RP. Studying a study and testing a test, 2nd edn. Boston: Little, Brown and Company, 1989.
- Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet* 2000; **355**: 129-34.
- Olsen O, Gotzsche PC. Cochrane review on screening for breast cancer with mammography. *Lancet* 2001; **358**: 1340-42.
- Horton R. Screening mammography: an overview revisited. *Lancet* 2001; **358**: 1284-85.

An overview of clinical research: the lay of the land

David A Grimes, Kenneth F Schulz

Many clinicians report that they cannot read the medical literature critically. To address this difficulty, we provide a primer of clinical research for clinicians and researchers alike. Clinical research falls into two general categories: experimental and observational, based on whether the investigator assigns the exposures or not. Experimental trials can also be subdivided into two: randomised and non-randomised. Observational studies can be either analytical or descriptive. Analytical studies feature a comparison (control) group, whereas descriptive studies do not. Within analytical studies, cohort studies track people forward in time from exposure to outcome. By contrast, case-control studies work in reverse, tracing back from outcome to exposure. Cross-sectional studies are like a snapshot, which measures both exposure and outcome at one time point. Descriptive studies, such as case-series reports, do not have a comparison group. Thus, in this type of study, investigators cannot examine associations, a fact often forgotten or ignored. Measures of association, such as relative risk or odds ratio, are the preferred way of expressing results of dichotomous outcomes—eg, sick versus healthy. Confidence intervals around these measures indicate the precision of these results. Measures of association with confidence intervals reveal the strength, direction, and a plausible range of an effect as well as the likelihood of chance occurrence. By contrast, p values address only chance. Testing null hypotheses at a p value of 0.05 has no basis in medicine and should be discouraged.

Clinicians today are in a bind. Increasing demands on their time are squeezing out opportunities to stay abreast of the literature, much less read it critically. Results of several studies indicate an inverse relation between knowledge of contemporary care and time since graduation from medical school.^{1,2} In many jurisdictions, attendance at a specified number of hours of continuing medical education courses is mandatory to maintain a licence to practise. However, the failure of these courses to improve patient care^{3,4} emphasises the importance of self-directed learning through reading. Many clinicians in practice, though, report that they feel unqualified to read the medical literature critically.⁵ Scientific illiteracy is a major failing of medical education.⁶

We have written this series of short essays on research methods for busy clinicians and active researchers. The needs of clinicians predominate; hopefully, this primer will produce more critical and thoughtful consumers of research, and thus better practitioners. The needs of clinicians overlap with those of researchers throughout the essays, but that overlap becomes most pronounced in the discussion of randomised controlled trials. For readers to assess randomised trials accurately, they should understand the relevant guidelines on the conduct of trials, emerging from methodological research. In presenting those discussions to clinicians, our essays will hopefully help researchers who do randomised trials as well.

We will cover descriptive studies, cohort studies, case-control studies, bias, and screening tests in separate articles, but will devote five articles to randomised controlled trials. This disproportion is intentional; randomised controlled trials are the gold standard in clinical research, and *The Lancet* publishes large numbers of them. Randomised controlled trials help to eliminate bias, and research has identified the important

methodological elements of trials that minimise bias.^{7,8} Finally, because trials are so important, clinicians might be more likely to act on their results than on those of observational studies; hence, investigators need to ensure that trials are done and reported well. Here, we provide a brief overview of research designs and discuss some of the common measures used.

A taxonomy of clinical research

Analogous to biological taxonomy, a simple hierarchy can be used to categorise most studies (panel).⁹ To do so, however, the study design must be known. As in biology, anatomy dictates physiology. The anatomy of a study determines what it can and cannot do. A difficulty that readers encounter is that authors sometimes do not report

Rating clinical evidence

Assessment system of the US Preventive Services Task Force

Quality of evidence

- I Evidence from at least one properly designed randomised controlled trial.
- II-1 Evidence obtained from well-designed controlled trials without randomisation.
- II-2 Evidence from well-designed cohort or case-control studies, preferably from more than one centre or research group.
- II-3 Evidence from multiple time series with or without the intervention. Important results in uncontrolled experiments (such as the introduction of penicillin treatment in the 1940s) could also be considered as this type of evidence.
- III Opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees.

Strength of recommendations

- A Good evidence to support the intervention.
- B Fair evidence to support the intervention.
- C Insufficient evidence to recommend for or against the intervention, but recommendation might be made on other grounds.
- D Fair evidence against the intervention.
- E Good evidence against the intervention.

Lancet 2002; **359**: 57–61

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (D A Grimes MD, K F Schulz PhD)

Correspondence to: Dr David A Grimes (e-mail: dgrimes@fhi.org)

the study type or provide sufficient detail to figure it out. A related problem is that authors sometimes incorrectly label the type of research done. Examples include calling non-randomised controlled trials randomised,¹⁰ and labelling non-concurrent cohort studies case-control studies.¹¹⁻¹³ The adjective case-controlled is also sometimes (inappropriately) applied to any study with a comparison group.

Biology has animal and plant kingdoms. Similarly, clinical research has two large kingdoms: experimental and observational research. Figure 1 shows that one can quickly decide the research kingdom by noting whether the investigators assigned the exposures—eg, treatments—or whether they observed usual clinical practice.¹⁴⁻¹⁸ For experimental studies, one needs to distinguish whether the exposures were assigned by a truly random technique (with concealment of the upcoming assignment from those involved) or whether some other allocation scheme was used, such as alternate assignment.¹⁹ An example of the latter would be a trial alternating months of liberal versus restricted access to electronic fetal monitoring for women in labour.²⁰

With observational studies, which dominate the literature,²¹ the next step is to ascertain whether the study has a comparison or control group. If so, the study is termed analytical. If not, it is a descriptive study (figure 1). If the study is analytical, the temporal direction of the trial needs to be identified. If the study determines both exposures and outcomes at one time point, it is termed cross-sectional. An example would be measurement of serum cholesterol of men admitted to a hospital with myocardial infarction versus that of their nextdoor neighbour. This type of study provides a snapshot of the population of sick and well at one time point.

If the study begins with an exposure—eg, oral contraceptive use—and follows women for a few years to

measure outcomes—eg, ovarian cancer—then it is deemed a cohort study. Cohort studies can be either concurrent or non-concurrent. By contrast, if the analytical study begins with an outcome—eg, ovarian cancer—and looks back in time for an exposure, such as use of oral contraceptives, then the study is a case-control study.

Studies without comparison groups are called descriptive studies. At the bottom of the research hierarchy is the case report.²² When more than one patient is described, it becomes a case-series report.²³

What studies can and cannot do

Is the study design appropriate for the question?

Starting at the bottom of the research hierarchy, descriptive studies are often the first foray into a new area of medicine. Investigators do descriptive studies to describe the frequency, natural history, and possible determinants of a condition.^{14,16,17} The results of these studies show how many people develop a disease or condition over time, describe the characteristics of the disease and those affected, and generate hypotheses about the cause of the disease. These hypotheses can be assessed through more rigorous research, such as analytical studies or randomised controlled trials. An example of a descriptive study would be the early reports of Legionnaire's disease²⁴ and toxic-shock syndrome.²⁵ An important caveat (often forgotten or intentionally ignored) is that descriptive studies, which do not have a comparison group, do not allow assessment of associations. Only comparative studies (both analytical and experimental) enable assessment of possible causal associations.

Cross-sectional study: a snapshot in time

Sometimes termed a frequency survey or a prevalence study,²⁶ cross-sectional studies are done to examine the presence or absence of disease and the presence or absence of an exposure at a particular time. Thus, prevalence, not incidence, is the focus. Since both outcome and exposure are ascertained at the same time (figure 2), the temporal relation between the two might be unclear. For example, assume that a cross-sectional study finds obesity to be more common among women with than without arthritis. Did the extra weight load on joints lead to arthritis, or did women with arthritis become involuntarily inactive and then obese? This type of question is unanswerable in a cross-sectional study.

Cohort study: looking forward in time

Cohort studies proceed in a logical sequence: from exposure to outcome (figure 2). Hence, this type of research is easier to understand than case-control studies. Investigators identify a group with an exposure of interest and another group or groups without the exposure. The investigators then follow the exposed and unexposed groups forward in time to determine outcomes. If the exposed group develops a higher incidence of the outcome than the unexposed, then the exposure is associated with an increased risk of the outcome.

The cohort study has important strengths and weaknesses. Because exposure is identified at the outset, one can assume that the exposure preceded the outcome. Recall bias is less of a concern than in the case-control study. The cohort study enables calculation of true incidence rates, relative risks, and attributable risks. However, for the study of rare events or events that take years to develop, this type of research design can be slow to yield results and thus prohibitively expensive. Nonetheless, several famous, large cohort studies²⁷⁻³⁰ continue to provide important information.

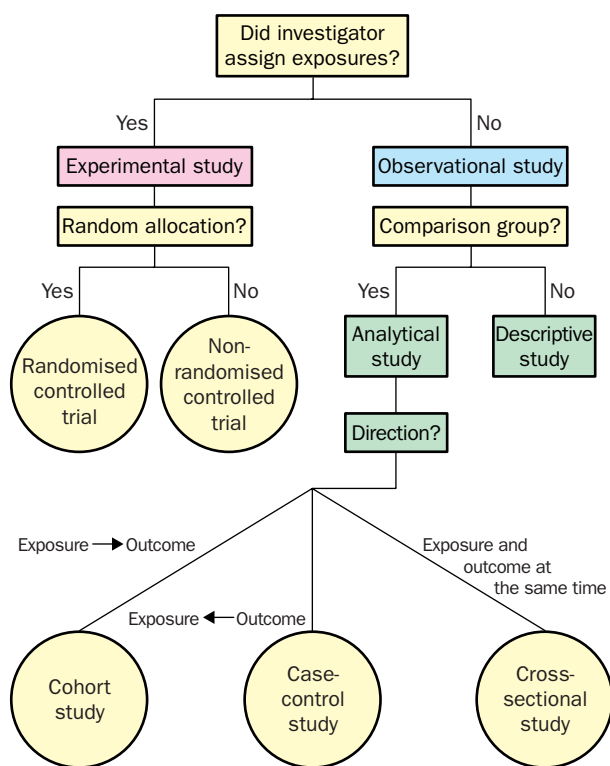


Figure 1: Algorithm for classification of types of clinical research

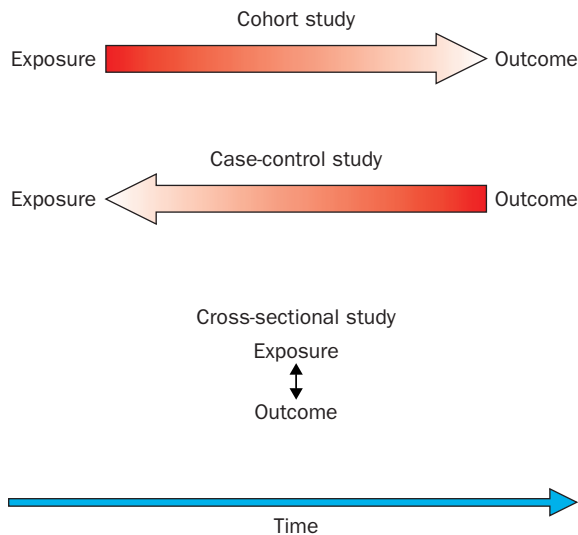


Figure 2: **Schematic diagram showing temporal direction of three study designs**

Case-control study: thinking backwards

Case-control studies work backwards. Because thinking in this direction is not intuitive for clinicians, case-control studies are widely misunderstood. Starting with an outcome, such as disease, this type of study looks backward in time for exposures that might have caused the outcome. As shown in figure 2, investigators define a group with an outcome (for example, ovarian cancer) and a group without the outcome (controls). Then, through chart reviews, interviews, or other means, the investigators ascertain the prevalence (or amount) of exposure to a risk factor—eg, oral contraceptives, ovulation-induction drugs—in both groups. If the prevalence of the exposure is higher among cases than among controls, then the exposure is associated with an increased risk of the outcome.

Case-control studies are especially useful for outcomes that are rare or that take a long time to develop, such as cardiovascular disease and cancer. These studies often require less time, effort, and money than would cohort studies. The Achilles heel of case-control studies is choosing an appropriate control group. Controls should be similar to cases in all important respects except for not having the outcome in question. Inappropriate control groups have ruined many case-control studies and caused much harm. Additionally, recall bias (better recollection of exposures among the cases than among the controls) is a persistent difficulty in studies that rely on memory. Because the case-control study lacks denominators, investigators cannot calculate incidence rates, relative risks, or attributable risks. Instead, odds ratios are the measure of association used; when the outcome is uncommon—eg, most cancers—the odds ratio provides a good proxy for the true relative risk.

Outbreaks of food-borne diseases are a prototype for case-control studies. On a cruise ship, the entire universe of those at risk is known. Those with vomiting and diarrhoea are asked about food exposures, as are a sample of those not ill. If a higher proportion of those ill reports having eaten a food than those well, the food becomes suspect. In this way, German potato salad on a ship was linked with a serious outbreak of shigella resistant to several antibiotics.³¹

Non-randomised trial: penultimate design?

Some experimental trials do not randomly allocate participants to exposures—eg, treatments or prevention

strategies. Instead of using truly random techniques, investigators often use methods that fall short of the mark—eg, alternate assignment.²⁰ The US Preventive Services Task Force⁹ and Canadian Task Force on the Periodic Health Examination³² designate this research design as class II-1, indicating less scientific rigour than randomised trials but more than analytical studies (panel).

After the investigators have assigned participants to treatment groups, the way a non-randomised trial is done and analysed resembles that of a cohort study. The exposed and unexposed are followed forward in time to ascertain the frequency of outcomes. Advantages of a non-randomised trial include use of a concurrent control group and uniform ascertainment of outcomes for both groups. However, selection bias can occur.

Randomised controlled trial: gold standard

The randomised controlled trial is the only known way to avoid selection and confounding biases in clinical research. This design approximates the controlled experiment of basic science. It resembles the cohort study in several respects, with the important exception of randomisation of participants to exposures (figure 2).

The hallmark of randomised controlled trials is assignment of participants to exposures purely by the play of chance. Randomised controlled trials reduce the likelihood of bias in determination of outcomes. When properly implemented, random allocation precludes selection bias. Trials feature uniform diagnostic criteria for outcomes and, often, blinding of those involved as to the exposure each participant is receiving, therefore reducing information bias. A unique strength of this study design is that it eliminates confounding bias, both known and unknown. Furthermore, the trial tends to be statistically efficient. If properly designed and done, a randomised controlled trial is likely to be free of bias and is thus especially useful for examination of small or moderate effects. In observational studies, bias might easily account for small to moderate differences.³³

Randomised controlled trials have drawbacks too. External validity is one. Whereas the randomised controlled trial, if properly done, has internal validity—ie, it measures what it sets out to measure—it might not have external validity. This term indicates the extent to which results can be generalised to the broader community. Unlike the observational study, the randomised controlled trial includes only volunteers who pass through a screening process before inclusion. Those who volunteer for trials tend to be different from those who do not; for example, their health might be better.³⁴ Another limitation is that a randomised controlled trial cannot be used in some instances, since intentional exposure to harmful substances—eg, toxins, bacteria, or other noxious exposures—would be unethical. As with cohort studies, the randomised controlled trial can be prohibitively expensive. Indeed, the cost of large trials runs into the tens of millions of US dollars.

Measurement of outcomes

Confusing fractions

Identification and quantification of outcomes is the business of research. However, slippery terminology often complicates matters for investigators and readers alike. For example, the term rate (as in maternal mortality rate) has been misused in textbooks and journal articles for decades. Additionally, rate is often used interchangeably with proportions and ratios.¹⁴ Figure 3 presents a simple approach to classification of these common terms.

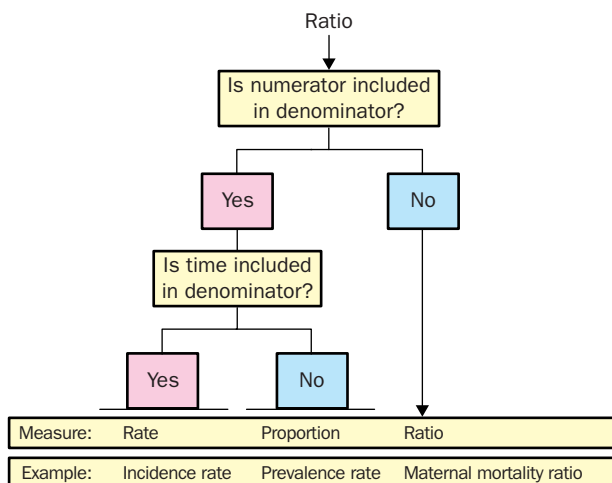


Figure 3: Algorithm for distinguishing rates, proportions, and ratios

A ratio is a value obtained by dividing one number by another.²⁶ These two numbers can be either related or unrelated. This feature—ie, relatedness of numerator and denominator—divides ratios into two groups: those in which the numerator is included in the denominator—eg, rate and proportion—and those in which it is not.

A rate measures the frequency of an event in a population. As shown in figure 3, the numerator (those with the outcome) of a rate must be contained in the denominator (those at risk of the outcome). Although all ratios feature a numerator and denominator, rates have two distinguishing characteristics: time and a multiplier. Rates indicate the time during which the outcomes occur and a multiplier, commonly to a base ten, to yield whole numbers. An example would be an incidence rate, indicating the number of new cases of disease in a population at risk over a defined interval of time—eg, 11 cases of tuberculosis per 100 000 persons per year.

Proportion is often used synonymously with rate, but the former does not have a time component. Like a rate, a proportion must have the numerator contained in the denominator.²⁶ Since the numerator and denominator have the same units, these divide out, leaving a dimensionless quantity; a number without units. An example of a proportion is prevalence—eg, 27 of 100 at risk have hay fever. This number indicates how many of a population at risk have a condition at a particular time (here, 27%); since documentation of new cases over time is not involved, prevalence is more properly considered a proportion than a rate.

Although all rates and proportions are ratios, the opposite is not true. In some ratios, the numerator is not included in the denominator. Perhaps the most notorious example is the maternal mortality ratio. The definition includes women who die of pregnancy-related causes in the numerator, and women with livebirths (usually 100 000) in the denominator. However, not all those in the numerator are included in the denominator—eg, a woman who dies of an ectopic pregnancy cannot be in the denominator of women with livebirths. Thus, this venerable misnomer is actually a ratio, not a rate, a fact only recently appreciated.

Measures of association: risky business

Relative risk (also termed the risk ratio)²⁶ is another useful ratio: the frequency of outcome in the exposed group divided by the frequency of outcome in the

unexposed. If the frequency of the outcome is the same in both groups, then the ratio is 1·0, indicating no association between exposure and outcome. By contrast, if the outcome is more frequent in those exposed, then the ratio will be greater than 1·0, implying an increased risk associated with exposure. Conversely, if the frequency of disease is less among the exposed, then the relative risk will be less than 1·0, implying a protective effect.

Also known as the cross-products ratio or relative odds,²⁶ the odds ratio has different meanings in different settings. In case-control studies, this measure is the usual measure of association. It indicates the odds of exposure among the case group divided by the odds of the exposure among controls. If cases and controls have equal odds of having the exposure, the odds ratio is 1·0, indicating no effect. If the cases have a higher odds of exposure than the controls, then the ratio is greater than 1·0, implying an increased risk associated with exposure. Similarly, odds ratios less than 1·0 indicate a protective effect.

An odds ratio can also be calculated for cross-sectional, cohort, and randomised controlled studies. Here, the disease-odds ratio is the ratio of the odds in favour of disease in the exposed versus that in the unexposed. In this context, the odds ratio has some appealing statistical features when studies are aggregated in meta-analysis, but the odds ratio does not indicate the relative risk when the proportion with the outcome is greater than 5–10%—ie, the term has little clinical relevance or meaning with higher incidence rates.³⁵

The confidence interval reflects the precision of study results. The interval provides a range of values for a variable, such as a proportion, relative risk, or odds ratio, that has a specified probability of containing the true value for the entire population from which the study sample was taken. Although 95% CIs are the most commonly used, others, such as 90%, are seen (and advocated).³⁶ The wider the confidence interval, the less precision exists in the result, and vice versa. For relative risks and odds ratios, when the 95% CI does not include 1·0, the difference is significant at the usual 0·05 level. However, use of this feature of confidence intervals as a back-door means of hypothesis testing is inappropriate.³⁶

Conclusion

Understanding what kind of study has been done is a prerequisite to thoughtful reading of research. Clinical research can be divided into experimental and observational; observational studies are further categorised into those with and without a comparison group. Only studies with comparison groups allow investigators to assess possible causal associations, a fact often forgotten or ignored. Dichotomous outcomes of studies should be reported as measures of association with confidence intervals; testing null hypotheses at arbitrary p values of 0·05 has no basis in medicine and should be discouraged.

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of this material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

- 1 Ramsey PG, Carline JD, Inui TS, et al. Changes over time in the knowledge base of practicing internists. *JAMA* 1991; **266**: 1103–07.
- 2 Evans CE, Haynes RB, Birkett NJ, et al. Does a mailed continuing education program improve physician performance? Results of a randomized trial in antihypertensive care. *JAMA* 1986; **255**: 501–04.

- 3 Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance: a systematic review of the effect of continuing medical education strategies. *JAMA* 1995; **274**: 700–05.
- 4 Sibley JC, Sackett DL, Neufeld V, Gerrard B, Rudnick KV, Fraser W. A randomized trial of continuing medical education. *N Engl J Med* 1982; **306**: 511–15.
- 5 Olatunbosun OA, Edouard L, Pierson RA. Physicians' attitudes toward evidence based obstetric practice: a questionnaire survey. *BMJ* 1998; **316**: 365–66.
- 6 Grimes DA, Bachicha JA, Learman LA. Teaching critical appraisal to medical students in obstetrics and gynecology. *Obstet Gynecol* 1998; **92**: 877–82.
- 7 Moher D, Schulz KF, Altman DG, Lepage L. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001; **357**: 1191–94.
- 8 Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; **134**: 663–94.
- 9 US Preventive Services Task Force. Guide to clinical preventive services, 2nd edn. Baltimore, MD: Williams & Wilkins, 1996.
- 10 Tatum HJ, Beltran RS, Ramos R, Van Kets H, Sivin I, Schmidt FH. Immediate postpartum insertion of GYNE-T 380 and GYNE-T 380 postpartum intrauterine contraceptive devices: randomized study. *Am J Obstet Gynecol* 1996; **175**: 1231–35.
- 11 Silver RK, Helfand BT, Russell TL, Ragin A, Sholl JS, MacGregor SN. Multifetal reduction increases the risk of preterm delivery and fetal growth restriction in twins: a case-control study. *Fertil Steril* 1997; **67**: 30–33.
- 12 Garg SK, Chase HP, Marshall G, Hoops SL, Holmes DL, Jackson WE. Oral contraceptives and renal and retinal complications in young women with insulin-dependent diabetes mellitus. *JAMA* 1994; **271**: 1099–102.
- 13 Berenson AB, Chacko MR, Wiemann CM, Mishaw CO, Friedrich WN, Grady JJ. A case-control study of anatomic changes resulting from sexual abuse. *Am J Obstet Gynecol* 2000; **182**: 820–34.
- 14 Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Company, 1987.
- 15 Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman RB, eds. Designing clinical research: an epidemiologic approach, 2nd edn. Baltimore: Lippincott Williams & Wilkins, 2001.
- 16 Kelsey JL, Whittemore AS, Evans AS, Thompson WD. Methods in observational epidemiology, 2nd edn. New York: Oxford University Press, 1996.
- 17 Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders Company, 1985.
- 18 Rothman KJ. Modern epidemiology. Boston: Little, Brown and Company, 1986.
- 19 Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA* 1994; **272**: 125–28.
- 20 Leveno KJ, Cunningham FG, Nelson S, et al. A prospective comparison of selective and universal electronic fetal monitoring in 34,995 pregnancies. *N Engl J Med* 1986; **315**: 615–19.
- 21 Funai EF, Rosenbush EJ, Lee MJ, Del Priore G. Distribution of study designs in four major US journals of obstetrics and gynecology. *Gynecol Obstet Invest* 2001; **51**: 8–11.
- 22 Nerlich AG, Zink A, Szeimies U, Hagedorn HG. Ancient Egyptian prosthesis of the big toe. *Lancet* 2000; **356**: 2176–79.
- 23 Zink A, Rohrbach H, Szeimies U, et al. Malignant tumors in an ancient Egyptian population. *Anticancer Res* 1999; **19**: 4273–77.
- 24 Keys TF. A sporadic case of pneumonia due to legionnaires disease. *Mayo Clin Proc* 1977; **52**: 657–60.
- 25 McKenna UG, Meadows JA 3rd, Brewer NS, Wilson WR, Perrault J. Toxic shock syndrome, a newly recognized disease entity. Report of 11 cases. *Mayo Clin Proc* 1980; **55**: 663–72.
- 26 Last JM, ed. A dictionary of epidemiology, 2nd edn. New York: Oxford University Press, 1988.
- 27 Huang Z, Willett WC, Colditz GA, et al. Waist circumference, waist: hip ratio, and risk of breast cancer in the Nurses' Health Study. *Am J Epidemiol* 1999; **150**: 1316–24.
- 28 Kim KS, Owen WL, Williams D, Adams-Campbell LL. A comparison between BMI and Conicity index on predicting coronary heart disease: the Framingham Heart Study. *Ann Epidemiol* 2000; **10**: 424–31.
- 29 Hannaford PC, Kay CR. The risk of serious illness among oral contraceptive users: evidence from the RCGP's oral contraceptive study. *Br J Gen Pract* 1998; **48**: 1657–62.
- 30 Doll R, Peto R, Boreham J, Sutherland I. Smoking and dementia in male British doctors: prospective study. *BMJ* 2000; **320**: 1097–102.
- 31 Lew JF, Swerdlow DL, Dance ME, et al. An outbreak of shigellosis aboard a cruise ship caused by a multiple- antibiotic-resistant strain of *Shigella flexneri*. *Am J Epidemiol* 1991; **134**: 413–20.
- 32 Canadian Task Force on the Periodic Health Examination. The Canadian guide to clinical preventive care. Ottawa: Minister of Supply and Services Canada, 1994.
- 33 MacMahon S, Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. *Lancet* 2001; **357**: 455–62.
- 34 Lilienfeld AM, Lilienfeld DE. Foundations of epidemiology, 2nd edn. New York: Oxford University Press, 1980.
- 35 Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! Evidence-based medicine 1996; **1**: 164–66.
- 36 Sterne JA, Smith GD. Sifting the evidence: what's wrong with significance tests? *BMJ* 2001; **322**: 226–31.

Descriptive studies: what they can and cannot do

David A Grimes, Kenneth F Schulz

Descriptive studies often represent the first scientific toe in the water in new areas of inquiry. A fundamental element of descriptive reporting is a clear, specific, and measurable definition of the disease or condition in question. Like newspapers, good descriptive reporting answers the five basic W questions: who, what, why, when, where . . . and a sixth: so what? Case reports, case-series reports, cross-sectional studies, and surveillance studies deal with individuals, whereas ecological correlational studies examine populations. The case report is the least-publishable unit in medical literature. Case-series reports aggregate individual cases in one publication. Clustering of unusual cases in a short period often heralds a new epidemic, as happened with AIDS. Cross-sectional (prevalence) studies describe the health of populations. Surveillance can be thought of as watchfulness over a community; feedback to those who need to know is an integral component of surveillance. Ecological correlational studies look for associations between exposures and outcomes in populations—eg, per capita cigarette sales and rates of coronary artery disease—rather than in individuals. Three important uses of descriptive studies include trend analysis, health-care planning, and hypothesis generation. A frequent error in reports of descriptive studies is overstepping the data: studies without a comparison group allow no inferences to be drawn about associations, causal or otherwise. Hypotheses about causation from descriptive studies are often tested in rigorous analytical studies.

Descriptive studies have several important roles in medical research. They are often the first foray into a new disease or area of inquiry—the first scientific “toe in the water”.¹ They document the health of populations and often prompt more rigorous studies. Since descriptive studies are often reported,² clinicians need to know their uses, strengths, and weaknesses.

A descriptive study is “concerned with and designed only to describe the existing distribution of variables, without regard to causal or other hypotheses.”³ The key qualifier about causal hypotheses is sometimes forgotten by investigators, resulting in erroneous conclusions. Here, we provide an overview of the advantages and disadvantages of descriptive studies, provide examples of several types of descriptive study, examine their clinical uses, and show how they can be misinterpreted.

The descriptive triad—or pentad?

Five “W” questions

Traditional descriptive epidemiology has focused on three key features: person, place, and time,⁴ or agent, host, and environment.⁵ An alternative approach is that of newspaper coverage. Good descriptive research, like good newspaper reporting, should answer five basic “W” questions—who, what, why, when, and where—and an implicit sixth question, so what?

Who has the disease in question? Age and sex are universally described, but other characteristics might be important too, including race, occupation, or recreational activities. The risk of venous thromboembolism, for example, increases exponentially with age.⁶ Only 1% of breast cancers arise in men, but Klinefelter’s syndrome or a family history of breast cancer increase their risk.^{7,8} Race affects the risk of leiomyomas of the uterus.⁹ Commercial fishing remains a risky business,¹⁰ and having fun with an

all-terrain vehicle¹¹ or snowmobile,¹² especially when drunk, can be lethal.

What is the condition or disease being studied? Development of a clear, specific, and measurable case definition is an essential step in descriptive epidemiology. Without such a description, the reader cannot interpret the report. Some conditions, such as fractures, can be overt. Other diagnoses might be challenging: multiple sclerosis, systemic lupus erythematosus, and pelvic inflammatory disease (salpingitis), for example. By use of the consensus or Delphi panel¹³ approach rather than evidence, some organisations have promulgated case definitions that have subsequently been shown to be invalid.¹⁴ For instance, evidence indicates that vaginal discharge and a raised erythrocyte sedimentation rate predict salpingitis,¹⁵ yet these predictors are not included in widely-used diagnostic criteria.¹⁴

Generally, stringent criteria for case definitions are desirable. Admittedly, if only the more severe cases of disease are targeted, milder or earlier cases will be missed. Although this approach inevitably leads to some loss of information, the trade-off is better specificity; severe cases of a disease are less likely to be confused with other conditions than are mild cases. An example would be the stringent case definition used for toxic shock syndrome, which requires involvement of multiple organ systems.¹⁶ More recently, expanding the case definition of AIDS has yielded a sudden surge in “new” cases.¹⁷

Why did the condition or disease arise? Descriptive studies often provide clues about cause that can be pursued with more sophisticated research designs (panel).

When is the condition common or rare? Time provides important clues about health events. The prototype might be the outbreak of gastroenteritis soon after ingestion of staphylococcal toxin. Some temporal relations can be long—eg, vaginal adenosis and clear cell carcinoma of the vagina appeared years after intrauterine exposure to diethylstilboestrol.¹⁸ Furthermore, cervical and other epithelial cancers develop decades after infection with human papillomavirus, and births and deaths from pneumonia and influenza have regular seasonal patterns, as might sperm counts.¹⁹

Lancet 2002; **359**: 145–49

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (D A Grimes MD, K F Schulz PhD)

Correspondence to: Dr David A Grimes
(e-mail: dgrimes@fhi.org)

Examples of early leads from descriptive studies

Clinical observation	Underlying association
Hepatocellular adenoma in young women	Exposure to high-dose oral contraceptives
Blindness in newborn infants	High ambient oxygen concentrations in incubators
Kaposi's sarcoma in young men	Infection with HIV-1
Angiosarcoma of the liver in employees	Industrial exposure to vinyl chloride
Cataracts, heart defects, and deafness in newborns	Maternal infection with rubella during pregnancy

Where does or does not the disease or condition arise? Geography has had a huge effect on health. Living close to rodents and insects (and thus their parasites) has shaped both medical and political history.²⁰ Living where drinking water has high fluoride protects against dental caries,²¹ whereas residing downwind from a lead smelter is less salutary.²² Latitude plays a part in both multiple sclerosis²³ and vitamin D deficiency;²⁴ sunlight might decrease²⁵ or increase²⁶ cancer risk.

So What? The implicit "W" relates to the public health effect. In view of the proliferation of descriptive reports,² what is their import? Is the condition a current and timely one? Is it serious? Are large numbers involved? Are its societal implications broad? Has it been studied before?²⁷ Although many descriptive reports herald new illnesses or monitor health, the net effect of others might be only thicker curricula vitae at the expense of thinner forests.

Types of descriptive studies

Descriptive studies consist of two major groups: those that deal with individuals and those that relate to populations. Studies that involve individuals are the case report, the case-series report, cross-sectional studies, and surveillance, whereas ecological correlational studies examine populations.⁴

Case report

The case report is the least publishable unit in the medical literature. Often, an observant clinician reports an unusual disease or association, which prompts further investigations with more rigorous study designs (panel). For example, a clinician, among others, reported benign hepatocellular adenomas, a rare tumour, in women who had taken oral contraceptives.²⁸ A large case-control study pursued this lead and confirmed a strong association between long-term use of high-dose pills and this rare, but sometimes deadly, tumour.²⁹ Not all case reports deal with serious health threats, however; some simply enliven the generally drab medical literature.^{30–32}

Case-series report

A case-series aggregates individual cases in one report. Sometimes, the appearance of several similar cases in a short period heralds an epidemic. For example, a cluster of homosexual men in Los Angeles with a similar clinical syndrome alerted the medical community to the AIDS epidemic in North America.³³ Whereas a report of a single unusual case might not trigger further investigation, a case-series of several unusual cases (in excess of what might be expected) adds to the concern. A convenient feature of case-series reports is that they can constitute the case group for a case-control study, which can then explore hunches about causes of disease.

Cross-sectional (prevalence) studies

Prevalence studies describe the health of populations. For example, in the USA, periodic surveys of the health status of the population are done by the federal government—eg, the Health Interview Survey and the Health and Nutrition Examination Survey. Analogous to the decennial census, these studies provide a snapshot of the population at a particular time.

Prevalence studies can be done in smaller populations as well. For example, the results of a survey done in a Puerto Rican pharmaceutical factory indicated an exceptionally high prevalence of gynaecomastia among employees (figure). This finding led to the hypothesis that exposure to ambient oestrogen dust in the plant might be the cause; serum concentrations of oestrogen lent support to the hypothesis. After improvements in dust control in the factory, the epidemic disappeared.³⁴ Similar prevalence studies have linked gynaecomastia with feeding of refugees³⁵ and tainted food.³⁶

Although generally distinguished from cohort and case-control studies, the cross-sectional study can be thought of as the case-control analogue of a population cohort study.³⁷ Since both exposure and outcome are ascertained at the same time (the defining feature of a cross-sectional study), costs are small and loss to follow-up is not a problem. However, because exposure and outcome are identified at one time point, the temporal sequence is often impossible to work out. An exception would be long-standing exposures, such as sex or blood type, which unquestionably preceded the outcome. For exposures that vary, information of aetiological relevance from the past might be more useful than current information.³⁷

Surveillance

Surveillance is another important type of descriptive study. Surveillance can be thought of as watchfulness over a community. A more formal definition is "the ongoing systematic collection, analysis, and interpretation of health data essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those who need to know."³⁸ The key feature here is feedback, as in a servomechanism. Prevention and control of the problem are fundamental parts of the feedback loop.

Surveillance can be either active or passive. Passive surveillance relies on data generally gathered through traditional channels, such as death certificates. By contrast, active surveillance searches for cases. The reporting of abortion-related deaths provides an example. Since 1972, the US Centers for Disease Control and Prevention has been doing active surveillance of these deaths in the USA, using multiple overlapping sources—ie, state maternal mortality study committees, professional organisations, newspapers, and colleagues in the specialty. By comparison with official statistics, active surveillance identifies about twice as many deaths.³⁹ Similarly, underreporting of maternal deaths remains an international problem.^{40–43}

Epidemiological surveillance has made important contributions to health, but none more impressive than smallpox eradication. Surveillance and containment were responsible for the elimination of smallpox from the world, an extraordinary public-health achievement.⁴⁴ Whereas mass immunisation of the world's population had failed, the approach of identification of cases through surveillance and then immunisation of susceptible persons in the surrounding communities stopped transmission. Without a non-human vector, the virus died out.



Photograph courtesy of Judson Van Wyk, MD, University of North Carolina School of Medicine, NC, USA.

Gynaecomastia, a condition associated with occupational exposure to oestrogen dust, feeding of refugees, and ingestion of tainted food

Ecological correlational studies

Correlational studies look for associations between exposures and outcomes in populations rather than in individuals.⁴ Because much data might already have been collected, correlational studies can be a convenient initial search for hypotheses. The measure of association between exposure and outcome is the correlation coefficient r , which indicates how linear is the relation between exposure and outcome. For example, death rates from coronary artery disease correlate with per capita sales of cigarettes.⁴⁵ By contrast, access to safe legal abortion is inversely correlated with maternal mortality.⁴⁶⁻⁴⁹ The range of potential associations to be explored is nearly limitless.⁵⁰

Correlational studies have important limitations—ie, the inability to link exposure to outcome in individuals and to control for confounding (a mixing or blurring of effects). An example of the latter is the observation that death rates from coronary artery disease also correlate with the number of colour television sets per capita.⁴ Even television's harshest critics are unlikely to argue that it clogs coronary vessels, an example of "ecological fallacy".³⁷ Although a link between television violence and violence in schools seems more plausible, whether this association is indeed causal is difficult to establish.⁵¹

Uses of descriptive studies

Trend analysis

Descriptive studies have several useful roles. Being able to monitor the health of populations is important to health-care administrators. Trend analysis is often provided by ongoing surveillance. Examples include the emerging epidemic of syphilis in the Russian Federation,^{52,53} and the international epidemic of multiple births, prematurity, and low birthweight caused by assisted reproductive technologies.⁵⁴⁻⁵⁸ Both epidemics raise troubling societal issues.

Planning

A second use is health-care planning. For example, the introduction of laparoscopy, coupled with bad press about oral contraceptives and intrauterine devices, tripled US rates of tubal sterilisation in the 1970s.⁵⁹ Hospitals and ambulatory surgery centres had a surge in demand for operations, yet less need for hospital beds. Similarly, the introduction of highly active antiretroviral therapy for patients with AIDS decreased bed occupancy.⁶⁰

Clues about cause

A third use of descriptive studies is to develop hypotheses about cause (panel). Observant clinicians noted an association between high concentrations of oxygen in incubators and blindness in babies; this finding led to analytical studies, then a randomised controlled trial, confirming the association.⁶¹ Unexpectedly high rates of cancer among women who had painted radium dials in watches alerted investigators to the danger of this occupational exposure.⁶²

Advantages and disadvantages

Descriptive studies have both strengths and weaknesses. Often, the data are already available and thus inexpensive and efficient to use. Furthermore, few ethical difficulties exist. However, descriptive studies have important limitations. Temporal associations between putative causes and effects might be unclear. A dangerous pitfall is that the investigators might draw causal inferences when none is possible.²⁷

Overstepping the data

A common mistake in inference is post hoc ergo propter hoc reasoning (after the thing, therefore on account of the thing), an example of a false cause.⁶³ In other words, a temporal association is incorrectly inferred to be a causal one. In one egregious example, seven women in Pasadena, California, created controversy around the world in the late 1980s. Seen in one physician's office, the women had developed functional ovarian cysts while taking the new multiphasic oral contraceptive pills.⁶⁴ Based on this uncontrolled observation, a case-series report warned that phasic pills might pose a threat to patient health and safety. The media printed the story, and unknown numbers of women around the world stopped taking their pills,⁶⁵ because they did not understand the difference between functional cysts and ovarian cancer. Since the report had no comparison group—eg, women using monophasic pills or those using none—the authors could not draw any conclusions about cause of disease.

In the wake of this report, damage-control efforts started quickly. Within 2 years, a publication showed no temporal association between the marketing of multiphasic pills and the number of women admitted to hospital for treatment of benign ovarian cysts.⁶⁶ However, 5 years elapsed before cohort⁶⁷ and case-control studies⁶⁸ confirmed no association between multiphasic pills and ovarian cysts. By this time, the public-health damage had been done.⁶⁹

Another sad example in which misinterpretation of descriptive studies hurt public health is routine electronic fetal monitoring in labour. A quarter of a century ago, temporal associations between the introduction of electronic fetal monitoring and falling perinatal mortality rates led to the conclusion that continuous fetal heart rate monitoring was a good thing.⁷⁰ Moreover, authorities of the day predicted a 50% reduction in perinatal morbidity and mortality from its use.⁷⁰

Based on this rosy assessment from prominent obstetricians, this expensive and intrusive technology took obstetrics by storm. However, the initial upbeat assessment did not survive scientific scrutiny. Years later, a meta-analysis of the randomised controlled trials showed that, by comparison with routine intermittent auscultation, routine electronic fetal monitoring confers no lasting benefit to infants, whereas it significantly increases operative deliveries; thus harming women.⁷¹ Based on objective reviews, both the Canadian Task Force on the Periodic Health Examination⁷² and the US

Preventive Services Task Force⁷³ have given routine electronic fetal monitoring a D recommendation (fair evidence against its routine use). Despite this advice, about three-fourths of all births in the USA include electronic fetal monitoring.⁷³ Failure to appreciate the limitations of descriptive studies has caused lasting harm and squandered billions of dollars.

Conclusion

Descriptive studies are often the first, tentative approach to a new event or condition. These studies generally emphasise features of a new disease or assess the health status of communities. Health administrators use descriptive studies to monitor trends and plan for resources. By contrast, epidemiologists and clinicians generally use descriptive reports to search for clues of cause of disease—ie, generation of hypotheses. In this role, descriptive studies are often a springboard into more rigorous studies with comparison groups. Common pitfalls of descriptive reports include an absence of a clear, specific, and reproducible case definition, and interpretations that overstep the data. Studies without a comparison group do not allow conclusions about cause of disease.

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of this material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

- Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman RB, eds. *Designing clinical research: an epidemiologic approach*, 2nd edn. Baltimore: Lippincott Williams and Wilkins, 2001.
- Fletcher RH, Fletcher SW. *Clinical research in general medical journals: a 30-year perspective*. *N Engl J Med* 1979; **301**: 180–83.
- Last JM, ed. *A dictionary of epidemiology*, 2nd edn. New York: Oxford University Press, 1988.
- Hennekens CH, Buring JE. *Epidemiology in medicine*. Boston: Little, Brown and Company, 1987.
- Lilienfeld AM, Lilienfeld DE. *Foundations of epidemiology*, 2nd edn. New York: Oxford University Press, 1980.
- Anderson FA Jr, Wheeler HB, Goldberg RJ, et al. A population-based perspective of the hospital incidence and case-fatality rates of deep vein thrombosis and pulmonary embolism. The Worcester DVT Study. *Arch Intern Med* 1991; **151**: 933–38.
- Sasco AJ, Lowenfels AB, Pasker-de Jong P. Review article: epidemiology of male breast cancer—a meta-analysis of published case-control studies and discussion of selected aetiological factors. *Int J Cancer* 1993; **53**: 538–49.
- Thomas DB. Breast cancer in men. *Epidemiol Rev* 1993; **15**: 220–31.
- Marshall LM, Spiegelman D, Barbieri RL, et al. Variation in the incidence of uterine leiomyoma among premenopausal women by age and race. *Obstet Gynecol* 1997; **90**: 967–73.
- Jaremin B, Kotulak E, Starnawska M, Mrozinski W, Wojciechowski E. Death at sea: certain factors responsible for occupational hazard in Polish seamen and deep-sea fishermen. *Int J Occup Med Environ Health* 1997; **10**: 405–16.
- Krane BD, Ricci MA, Sweeney WB, Deshmukh N. All-terrain vehicle injuries: a review at a rural level II trauma center. *Am Surg* 1988; **54**: 471–74.
- Gabert T, Stueland DT. Recreational injuries and deaths in northern Wisconsin: analysis of injuries and fatalities from snowmobiles over 3 years. *Wis Med J* 1993; **92**: 671–75.
- Zinn J, Zalokowski A. The use of the Delphi panel for consensus development on indicators of laboratory performance. *Clin Lab Manage Rev* 1999; **13**: 386–408.
- Hager WD, Eschenbach DA, Spence MR, Sweet RL. Criteria for diagnosis and grading of salpingitis. *Obstet Gynecol* 1983; **61**: 113–14.
- Hadgu A, Westrom L, Brooks CA, Reynolds GH, Thompson SE. Predicting acute pelvic inflammatory disease: a multivariate analysis. *Am J Obstet Gynecol* 1986; **155**: 954–60.
- Wiesenthal AM, Ressler M, Caston SA, Todd JK. Toxic shock syndrome, I: clinical exclusion of other syndromes by strict and screening definitions. *Am J Epidemiol* 1985; **122**: 847–56.
- Anon. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *MMWR Morb Mortal Wkly Rep* 1992; **41**: 1–19.
- Yon JL Jr, Lutz MH, Girtanner RE, Averette HE. Adenosis and adenocarcinoma of the vagina in young women: a review. *Gynecol Oncol* 1974; **2**: 508–17.
- Levine RJ, Mathew RM, Chenault CB, et al. Differences in the quality of semen in outdoor workers during summer and winter. *N Engl J Med* 1990; **323**: 12–16.
- McNeill WH. *Plagues and peoples*. New York: Anchor Books, 1998.
- Parko A. Longitudinal study of dental caries prevalence and incidence in the rapakivi (high fluoride) and olivine diabase (low fluoride) areas of Laitila, Finland. *Proc Finn Dent Soc* 1990; **86**: 103–06.
- Diaz-Barriga F, Batres L, Calderon J, et al. The El Paso smelter 20 years later: residual impact on Mexican children. *Environ Res* 1997; **74**: 11–16.
- Ebers GC, Sadovnick AD. The geographic distribution of multiple sclerosis: a review. *Neuroepidemiology* 1993; **12**: 1–5.
- Chapuy MC, Preziosi P, Maamer M, et al. Prevalence of vitamin D insufficiency in an adult normal population. *Osteoporos Int* 1997; **7**: 439–43.
- Ainsleigh HG. Beneficial effects of sun exposure on cancer mortality. *Prev Med* 1993; **22**: 132–40.
- Fears TR, Scotto J, Schneiderman MA. Skin cancer, melanoma, and sunlight. *Am J Public Health* 1976; **66**: 461–64.
- Wingo PA, Higgins JE, Rubin GL, Zahniser SC, eds. *An epidemiologic approach to reproductive health*. Geneva: WHO, 1994.
- Schenken JR. Hepatocellular adenoma: relationship to oral contraceptives? *JAMA* 1976; **236**: 559.
- Rooks JB, Ory HW, Ishak KG, et al. Epidemiology of hepatocellular adenoma: the role of oral contraceptive use. *JAMA* 1979; **242**: 644–48.
- Waugh RJ. Penile frostbite, an unforeseen hazard of jogging. *N Engl J Med* 1977; **296**: 178.
- Levit F. Jogger's nipples. *N Engl J Med* 1977; **297**: 1127.
- McBride DQ, Lehman LP, Mangiardi JR. Break-dancing neck. *N Engl J Med* 1985; **312**: 186.
- Anon. Pneumocystis pneumonia: Los Angeles. *MMWR Morb Mortal Wkly Rep* 1981; **30**: 250–52.
- Harrington JM, Stein GF, Rivera RO, de Morales AV. The occupational hazards of formulating oral contraceptives: a survey of plant employees. *Arch Environ Health* 1978; **33**: 12–15.
- Sattin RW, Roisin A, Kafriksen ME, Dugan JB, Farer LS. Epidemic of gynecomastia among illegal Haitian entrants. *Public Health Rep* 1984; **99**: 504–10.
- Fara GM, Del Corvo G, Bernuzzi S, et al. Epidemic of breast enlargement in an Italian school. *Lancet* 1979; **2**: 295–97.
- Rothman KJ. *Modern epidemiology*. Boston: Little, Brown and Company, 1986.
- Centers for Disease Control. *Comprehensive plan for epidemiologic surveillance*. Atlanta, GA: Centers for Disease Control, 1986.
- Cates W Jr, Smith JC, Rochat RW, Patterson JE, Dolman A. Assessment of surveillance and vital statistics data for monitoring abortion mortality, United States, 1972–1975. *Am J Epidemiol* 1978; **108**: 200–06.
- Kao S, Chen LM, Shi L, Weinrich MC. Underreporting and misclassification of maternal mortality in Taiwan. *Acta Obstet Gynecol Scand* 1997; **76**: 629–36.
- Schuitmaker N, Van Roosmalen J, Dekker G, Van Dongen P, Van Geijn H, Gravenhorst JB. Underreporting of maternal mortality in The Netherlands. *Obstet Gynecol* 1997; **90**: 78–82.
- Bouvier-Colle MH, Varnoux N, Costes P, Hatton F. Reasons for the underreporting of maternal mortality in France, as indicated by a survey of all deaths among women of childbearing age. *Int J Epidemiol* 1991; **20**: 717–21.
- Comas A, Navarro A, Conde J, Blasini I, Adamsons K. Misreporting of maternal mortality in Puerto Rico. *Bol Asoc Med P R* 1990; **82**: 343–46.
- Foege WH. Smallpox eradication in west and central Africa revisited. *Bull World Health Organ* 1998; **76**: 233–35.
- Friedman GD. Cigarette smoking and geographic variation in coronary heart disease mortality in the United States. *J Chronic Dis* 1967; **20**: 769–79.
- Anon. Abortion: one Romania is enough. *Lancet* 1995; **345**: 137–38.
- Stephenson P, Wagner M, Badea M, Serbanescu F. The public health consequences of restricted induced abortion: lessons from Romania. *Am J Public Health* 1992; **82**: 1328–31.
- Seward PN, Ballard CA, Ulene AL. The effect of legal abortion on the rate of septic abortion at a large county hospital. *Am J Obstet Gynecol* 1973; **115**: 335–38.
- Stewart GK, Goldstein PJ. Therapeutic abortion in California: effects of septic abortion and maternal mortality. *Obstet Gynecol* 1971; **37**: 510–14.

- 50 Noller KL, Resseguie LJ, Voss V. The effect of changes in atmospheric pressure on the occurrence of the spontaneous onset of labor in term pregnancies. *Am J Obstet Gynecol* 1996; **174**: 1192–99.
- 51 Centerwall BS. Television and violence: the scale of the problem and where to go from here. *JAMA* 1992; **267**: 3059–63.
- 52 Borisenko KK, Tichonova LI, Renton AM. Syphilis and other sexually transmitted infections in the Russian Federation. *Int J STD AIDS* 1999; **10**: 665–68.
- 53 Tichonova L, Borisenko K, Ward H, Meheus A, Gromyko A, Renton A. Epidemics of syphilis in the Russian Federation: trends, origins, and priorities for control. *Lancet* 1997; **350**: 210–13.
- 54 Tough SC, Greene CA, Svenson LW, Belik J. Effects of in vitro fertilization on low birth weight, preterm delivery, and multiple birth. *J Pediatr* 2000; **136**: 618–22.
- 55 Bider D, Livshitz A, Tur Kaspa I, Shulman A, Levron J, Dor J. Incidence and perinatal outcome of multiple pregnancies after intracytoplasmic sperm injection compared to standard in vitro fertilization. *J Assist Reprod Genet* 1999; **16**: 221–26.
- 56 Steegers-Theunissen RP, Zwertbroek WM, Huisjes AJ, Kanhai HH, Bruinse HW, Merkus HM. Multiple birth prevalence in The Netherlands: impact of maternal age and assisted reproductive techniques. *J Reprod Med* 1998; **43**: 173–79.
- 57 Dunn A, Macfarlane A. Recent trends in the incidence of multiple births and associated mortality in England and Wales. *Arch Dis Child Fetal Neonatal Ed* 1996; **75**: F10–19.
- 58 Anon. Pregnancies and births resulting from in vitro fertilization: French national registry, analysis of data 1986 to 1990. FIVNAT (French In Vitro National). *Fertil Steril* 1995; **64**: 746–56.
- 59 Peterson HB, Greenspan JR, DeStefano F, Ory HW, Layde PM. The impact of laparoscopy on tubal sterilization in United States hospitals, 1970 and 1975 to 1978. *Am J Obstet Gynecol* 1981; **140**: 811–14.
- 60 Anon. Update: trends in AIDS incidence, deaths, and prevalence—United States, 1996. *MMWR Morb Mortal Wkly Rep* 1997; **46**: 165–73.
- 61 Silverman WA. Memories of the 1953–54 Oxygen Trial and its aftermath: the failure of success. *Control Clin Trials* 1991; **12**: 355–58.
- 62 Stebbings JH, Lucas HF, Stehney AF. Mortality from cancers of major sites in female radium dial workers. *Am J Ind Med* 1984; **5**: 435–59.
- 63 Copi IM, Cohen C. Introduction to logic, 10th edn. Upper Saddle River, NJ: Prentice Hall, 1998.
- 64 Caillouette JC, Koehler AL. Phasic contraceptive pills and functional ovarian cysts. *Am J Obstet Gynecol* 1987; **156**: 1538–42.
- 65 Martinez F. Responsibility of health providers and the media in response to scientific information. *Eur J Contracept Reprod Health Care* 1997; **2**: 25–30.
- 66 Grimes DA, Hughes JM. Use of multiphasic oral contraceptives and hospitalizations of women with functional ovarian cysts in the United States. *Obstet Gynecol* 1989; **73**: 1037–39.
- 67 Lanes SF, Birmann B, Walker AM, Singer S. Oral contraceptive type and functional ovarian cysts. *Am J Obstet Gynecol* 1992; **166**: 956–61.
- 68 Holt VL, Daling JR, McKnight B, Moore D, Stergachis A, Weiss NS. Functional ovarian cysts in relation to the use of monophasic and triphasic oral contraceptives. *Obstet Gynecol* 1992; **79**: 529–33.
- 69 Jones EF, Beniger JR, Westoff CF. Pill and IUD discontinuation in the United States, 1970–1975: the influence of the media. *Fam Plann Perspect* 1980; **12**: 293–300.
- 70 Quilligan EJ, Paul RH. Fetal monitoring: is it worth it? *Obstet Gynecol* 1975; **45**: 96–100.
- 71 Thacker SB, Stroup DF, Peterson HB. Efficacy and safety of intrapartum electronic fetal monitoring: an update. *Obstet Gynecol* 1995; **86**: 613–20.
- 72 Canadian Task Force on the Periodic Health Examination. The Canadian guide to clinical preventive care. Ottawa: Minister of Supply and Services Canada, 1994.
- 73 US Preventive Services Task Force. Guide to clinical preventive services, 2nd edn. Baltimore, MD: Williams and Wilkins, 1996.

Bias and causal associations in observational research

David A Grimes, Kenneth F Schulz

Readers of medical literature need to consider two types of validity, internal and external. Internal validity means that the study measured what it set out to; external validity is the ability to generalise from the study to the reader's patients. With respect to internal validity, selection bias, information bias, and confounding are present to some degree in all observational research. Selection bias stems from an absence of comparability between groups being studied. Information bias results from incorrect determination of exposure, outcome, or both. The effect of information bias depends on its type. If information is gathered differently for one group than for another, bias results. By contrast, non-differential misclassification tends to obscure real differences. Confounding is a mixing or blurring of effects: a researcher attempts to relate an exposure to an outcome but actually measures the effect of a third factor (the confounding variable). Confounding can be controlled in several ways: restriction, matching, stratification, and more sophisticated multivariate techniques. If a reader cannot explain away study results on the basis of selection, information, or confounding bias, then chance might be another explanation. Chance should be examined last, however, since these biases can account for highly significant, though bogus results. Differentiation between spurious, indirect, and causal associations can be difficult. Criteria such as temporal sequence, strength and consistency of an association, and evidence of a dose-response effect lend support to a causal link.

Clinicians face two important questions as they read medical research: is the report believable, and, if so, is it relevant to my practice? Uncritical acceptance of published research has led to serious errors and squandered resources.¹ Here, we will frame these two questions in terms of study validity, describe a simple checklist for readers, and offer some criteria by which to judge reported associations.

Internal and external validity

Analogous to a laboratory test, a study should have internal validity—ie, the ability to measure what it sets out to measure.² The inference from participants in a study should be accurate. In other words, a research study should avoid bias or systematic error.³ Internal validity is the sine qua non of clinical research; extrapolation of invalid results to the broader population is not only worthless but potentially dangerous.

A second important concern is external validity; can results from study participants be extrapolated to the reader's patients? Since a total enumeration or census approach to medical research is usually impossible, the customary tactic is to choose a sample, study it, and, hopefully, extrapolate the result to one's practice. Gauging external validity is necessarily more subjective than is assessment of internal validity.

Internal and external validity entail important trade-offs. For example, randomised controlled trials are more likely than observational studies to be free of bias,⁴ but, because they usually enrol selected participants, external validity can suffer. This problem of unsuitable participants is also termed distorted assembly.⁵ Participants in randomised controlled trials tend to be different (including being healthier⁶⁻⁸) from those who choose not to take part, a function of the restricted entry

criteria. The filtering process for admission to randomised trials might, therefore, result in "a type of hothouse flower, which cannot bloom or be successfully removed beyond its special greenery."⁵

Bias

Bias undermines the internal validity of research. Unlike the conventional meaning of bias—ie, prejudice—bias in research denotes deviation from the truth. All observational studies (and, regrettably, many badly done randomised controlled trials)^{9,10} have built-in bias; the challenge for investigators, editors, and readers is to ferret these out and judge how they might have affected results. A simple checklist, such as that shown in panel 1, can be helpful.¹¹⁻¹⁴

Several taxonomies exist for classification of biases in clinical research. Sackett's landmark compilation,¹⁵ for example, included 35 different biases. By contrast Feinstein⁹ consolidated biases into four categories that arise sequentially during research: susceptibility, performance, detection, and transfer. Susceptibility bias refers to differences in baseline characteristics, performance bias to different proficiencies of treatment, detection bias to different measurement of outcomes, and transfer bias to differential losses to follow-up. Another approach,^{3,11,16,17} which is often used, is to group all biases into three general categories: selection, information, and confounding. The leitmotif for all three is "different".¹⁷ Something "different" distorts the planned comparison.

Selection bias

Are the groups similar in all important respects?

Selection bias stems from an absence of comparability between groups being studied. For example, in a cohort study, the exposed and unexposed groups differ in some important respect aside from the exposure. Membership bias is a type of selection bias: people who choose to be members of a group—eg, joggers—might differ in important respects from others. For instance, both cohort and case-control studies initially suggested that jogging after myocardial infarction prevented repeat

Lancet 2002; **359**: 248–52

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (D A Grimes MD, K F Schulz PhD)

Correspondence to: Dr David A Grimes
(e-mail: dgrimes@fhi.org)

Panel 1: What to look for in observational studies**Is selection bias present?**

In a cohort study, are participants in the exposed and unexposed groups similar in all important respects except for the exposure?

In a case-control study, are cases and controls similar in all important respects except for the disease in question?

Is information bias present?

In a cohort study, is information about outcome obtained in the same way for those exposed and unexposed?

In a case-control study, is information about exposure gathered in the same way for cases and controls?

Is confounding present?

Could the results be accounted for by the presence of a factor—eg, age, smoking, sexual behaviour, diet—associated with both the exposure and the outcome but not directly involved in the causal pathway?

If the results cannot be explained by these three biases, could they be the result of chance?

What are the relative risk or odds ratio and 95% CI?^{11,12}

Is the difference statistically significant, and, if not, did the study have adequate power to find a clinically important difference?^{13,14}

If the results still cannot be explained away, then (and only then) might the findings be real and worthy of note.

infarction. However, a randomised controlled trial failed to confirm this benefit.¹⁵ Those who chose to exercise might have differed in other important ways from those who did not exercise, such as diet, smoking, and presence of angina.

In case-control studies, selection bias implies that cases and controls differ importantly aside from the disease in question. Two types of selection bias have earned eponyms: Berkson and Neyman bias. Also known as an admission-rate bias, Berkson bias (or paradox) results from differential rates of hospital admission for cases and controls. Berkson initially thought that this phenomenon was due to presence of a simultaneous disease.⁵ Alternatively, knowledge of the exposure of interest might lead to an increased rate of admission to hospital. For example, doctors who care for women with salpingitis were more likely to recommend hospital admission for those using an intrauterine device (IUD) than for those using a hormonal method of contraception.^{18,19} In a hospital-based case-control study, this would stack the deck (or gynaecology ward) with a high proportion of IUD-exposed cases, spuriously increasing the odds ratio.

Neyman bias is an incidence-prevalence bias. It arises when a gap in time occurs between exposure and selection of study participants. This bias crops up in studies of diseases that are quickly fatal, transient, or subclinical. Neyman bias creates a case group not representative of cases in the community. For example, a hospital-based case-control study of myocardial infarction and snow shovelling (the exposure of interest) would miss individuals who died in their driveways and thus never reached a hospital; this eventuality might greatly lower the odds ratio of infarction associated with this strenuous activity.

Other types of selection bias include unmasking (detection signal) and non-respondent bias. An exposure might lead to a search for an outcome, as well as the outcome itself. For example, oestrogen replacement

therapy might cause symptomless endometrial cancer patients to bleed, resulting in initiation of diagnostic tests.²⁰ In this instance, the exposure unmasked the subclinical cancer, leading to a spurious increase in the odds ratio. In observational studies, non-respondents are different from respondents. Cigarette smokers are a case in point: smokers are less likely to return questionnaires than are non-smokers or pipe and cigar smokers.²¹

Information bias

Has information been gathered in the same way?

Information bias, also known as observation, classification, or measurement bias, results from incorrect determination of exposure or outcome, or both. In a cohort study or randomised controlled trial, information about outcomes should be obtained the same way for those exposed and unexposed. In a case-control study, information about exposure should be gathered in the same way for cases and controls.

Information bias can arise in many ways. Some use the term ascertainment to describe gathering information in different ways. For example, an investigator might gather information about an exposure at bedside for a case but by telephone from a community control. Diagnostic suspicion bias implies that knowledge of a putative cause of disease might launch a more intensive search for the disease among those exposed, for example, preferentially searching for infection by HIV-1 in intravenous drug users. Conversely, the presence of a disease might prompt a search for the putative exposure of interest. Another type of bias is family history bias, in which medical information flows differently to affected and unaffected family members, as has been shown for rheumatoid arthritis.²² To minimise information bias, detail about exposures in case-control studies should be gathered by people who are unaware of whether the respondent is a case or a control. Similarly, in a cohort study with subjective outcomes, the observer should be unaware of the exposure status of each participant.

In case-control studies that rely on memory of remote exposures, recall bias is pervasive. Cases tend to search their memories to identify what might have caused their disease; healthy controls have no such motivation. Thus, better recall among cases is common. For example, the putative association between induced abortion and subsequent development of breast cancer has emerged as a hot medical and political issue. Many case-control studies have reported an increase in cancer risk after abortion.²³ However, when investigators compared histories of prior abortions, obtained by personal interview, against centralised medical records, they documented systematic underreporting of abortions among controls (but not among cases) that accounted for a spurious association.²⁴ In Swedish and Danish cohort studies,^{25,26} free from recall bias, induced abortion has had either a protective effect or no effect on risk of breast cancer.

Is the information bias random or in one direction?

The effect of information bias depends on its type. If information is gathered differentially for one group than for another, then bias results, raising or lowering the relative risk or odds ratio dependent on the direction of the bias. By contrast, non-differential misclassification—ie, noise in the system—tends to obscure real differences. For example, an ambiguous questionnaire might lead to errors in data collection among cases and controls, shifting the odds ratio toward unity, meaning no association.

Confounding

Is an extraneous factor blurring the effect?

Confounding is a mixing or blurring of effects. A researcher attempts to relate an exposure to an outcome, but actually measures the effect of a third factor, termed a confounding variable. A confounding variable is associated with the exposure and it affects the outcome, but it is not an intermediate link in the chain of causation between exposure and outcome.^{27,28} More simply, confounding is a methodological fly in the ointment. Confounding is often easier to understand from examples than from definitions.

Oral contraceptives and myocardial infarction, and smoking

Early studies of the safety of oral contraceptives reported a pronounced increased risk of myocardial infarction. This association later proved to be spurious, because of the high proportion of cigarette smokers among users of birth control pills.²⁹⁻³¹ Here, cigarette smoking confounded the relation between oral contraceptives and infarction. Women who chose to use birth control pills also chose, in large numbers, to smoke cigarettes, and cigarettes, in turn, increased the risk of myocardial infarction. Although investigators thought they were measuring an effect of birth control pills, they were in fact measuring the hidden effect of smoking among pill users.

IUD insertion and salpingitis, and exposure to sexually transmitted disease

Results of a large case-control study of IUDs indicated a significant increase in salpingitis soon after insertion.³² However, among married or cohabiting women with only one reported sex partner in the past 6 months, no significant increase in risk was evident.³³ In the study, exposure to sexually transmitted diseases apparently confounded the association. Even among women at low risk of salpingitis, frequent coitus might increase risk of infection,³⁴ and few studies have controlled for this variable.

Oral contraceptives and cervical cancer, and smoking

Reported associations between oral contraceptives and squamous cervical cancer³⁵ might be due to unsuspected confounding by cigarette smoking and human papillomavirus infection.³⁶ Control of confounding is inevitably limited by our meagre understanding of human biology; unsuspected confounding factors evade control in observational studies.³⁷

Control for confounding

When selection bias or information bias exist in a study, irreparable damage results. Internal validity is doomed. By contrast, when confounding is present, this bias can be corrected, provided that confounding was anticipated and the requisite information gathered. Confounding can be controlled for before or after a study is done. The purpose of these approaches is to achieve homogeneity between study groups.

Restriction

The simplest approach is restriction (also called exclusion or specification).²⁸ For example, if cigarette smoking is suspected to be a confounding factor, a study can enrol only non-smokers. Although this tactic avoids confounding, it also hinders recruitment (and thus power) and precludes extrapolation to smokers. Restriction might increase the internal validity of a study at the cost of poorer external validity.

Matching

Another way to control for confounding is pairwise matching. In a case-control study in which smoking is deemed a confounding factor, cases and controls can be matched by smoking status. For each case who smokes, a control who smokes is found. This approach, although often used by investigators, has two drawbacks. If matching is done on several potential confounding factors, the recruitment process can be cumbersome, and, by definition, one cannot examine the effect of a matched variable.²⁸

Stratification

Investigators can also control for confounding after a study has been completed. One approach is stratification. Stratification can be considered a form of post hoc restriction, done during the analysis rather than during the accrual phase of a study. For example, results can be stratified by levels of the confounding factor. In the smoking example, results are calculated separately for smokers and non-smokers to see if the same effect arises independent of smoking. The Mantel-Haenszel procedure³⁸ combines the various strata into a summary statistic that describes the effect. The strata are weighted inversely to their variance—ie, strata with larger numbers count more than those with smaller numbers. If the Mantel-Haenszel adjusted effect differs substantially from the crude effect, then confounding is deemed present. In this instance, the adjusted estimate of effect is considered the better estimate to use.

Confounding is not always intuitive, as shown by the fictitious example in the figure. In this hypothetical

		Salpingitis		Total	Proportion with salpingitis
		Yes	No		
All women (n=2000)	Use of IUD				
	Yes	45	955	1000	4.5%
	No	15	985	1000	1.5%

Crude RR = $\frac{4.5\%}{1.5\%} = 3.0$ (95% CI 1.7-5.4)

		Salpingitis		Total	Proportion with salpingitis
		Yes	No		
Women with 1 sexual partner (n=1200)	Use of IUD				
	Yes	3	297	300	1.0%
	No	9	891	900	1.0%

RR = $\frac{1.0\%}{1.0\%} = 1.0$

		Salpingitis		Total	Proportion with salpingitis
		Yes	No		
Women with >1 sexual partner (n=800)	Use of IUD				
	Yes	42	658	700	6.0%
	No	6	94	100	6.0%

RR = $\frac{6.0\%}{6.0\%} = 1.0$

Example of confounding in a hypothetical cohort study of intrauterine device use and salpingitis

When the crude relative risk is controlled for the confounding effect of number of sexual partners, the raised risk disappears.

cohort of 2000 women, use of an IUD was strongly related to development of salpingitis (relative risk 3.0; 95% CI 1.7–5.4). However, the number of sexual partners was related to women's choice of contraception and to their risk of upper-genital-tract infection. Here, a disproportionate number of women with more than one sexual partner chose to use an IUD (700 *vs* 300 women with only one partner). The number of partners was also related to the risk of infection (6% among those with >1 partner *vs* 1% among those with only one partner). In each stratum by number of partners, the relative risk is 1.0, indicating no association between the IUD and salpingitis. The Mantel-Haenszel weighted relative risk, which controls for this confounding effect, is 1.0 (95% CI 0.5–2.0). In this fictitious example, the apparent three-fold increase in risk associated with IUD use was all due to confounding bias.

Multivariate techniques

In multivariate techniques, mathematical modelling examines the potential effect of one variable while simultaneously controlling for the effect of many other factors. A major advantage of these approaches is that they can control for more factors than can stratification. For example, an investigator might use multivariate logistic regression to study the effect of oral contraceptives on ovarian cancer risk. In this way, they could simultaneously control for age, race, family history, parity, &c. Another example would be use of a proportional hazards regression analysis for time to death; this method could control simultaneously for age, blood pressure, smoking history, serum lipids, and other risk factors.³⁹ Disadvantages of multivariate approaches, for some researchers, include greater difficulty in understanding the results, and loss of hands-on feel for the data.²⁸

Chance

If a reader cannot explain results on the basis of selection, information, or confounding bias, then chance might be another explanation. The reason for examination of bias before chance is that biases can easily cause highly significant (though bogus) results. Regrettably, many readers use the *p* value as the arbiter of validity, without considering these other, more important, factors.

The venerable *p* value measures chance. It advises the reader of the likelihood of a false-positive conclusion: a difference was seen in the study, although it does not exist in the broader population (type I error). Many clinicians are surprised to learn, however, that the *p* value of 0.05 as a threshold has no basis in medicine. Rather, it stems from agricultural and industrial experiments early in the 20th century.^{40,41} Should a study not achieve significance at this level, one needs to see if the study had adequate power to find a clinically important difference. Many “negative” studies simply have too few participants to do the job.^{13,14} Better yet, investigators should present measures of association with confidence intervals⁴¹ in preference to hypothesis tests.

Judgment of associations

Bogus, indirect, or real?

When statistical associations emerge from clinical research, the next step is to judge what type of association exists. Statistical associations do not necessarily imply causal associations.¹⁷ Although several classifications are available,²⁸ a simple approach includes just three types: spurious, indirect, and causal. Spurious associations are the result of selection bias, information bias, and chance.

By contrast, indirect associations (which stem from confounding) are real but not causal.

Judgment of cause-effect relations can be tough. Few rules apply, though criteria first suggested by Hill have received the most attention (panel 2).^{17,42,43} The only iron-clad criterion is temporality: the cause must antedate the effect. However, in many studies, especially with chronic diseases, answering this chicken-egg question can be daunting. Strong associations argue for causation. Whereas weak associations in observational studies can easily be due to bias, large amounts of bias would be necessary to produce strong associations. (This large bias is evident in reports that link IUD use with salpingitis.) Some suggest that relative risks more than 3 in cohort studies, or odds ratios greater than 4 in case-control studies, provide strong support for causation.⁴⁴ Consistent observation of an association in different populations and with different study designs also lends support to a real effect. For example, results of studies done around the world have consistently shown that oral contraceptives protect against ovarian cancer; a causal relation can, therefore, be argued. Evidence of a biological gradient supports a causal association too. For instance, protection against ovarian cancer is directly related to duration of use of oral contraceptives.⁴⁵ The risk of death from lung cancer is linearly related to years of cigarette smoking. In both of these examples, increasing exposure is associated with an increasing biological effect.

Other criteria of Hill's are less useful. Specificity is a weak criterion. With a few exceptions, such as the rabies virus, few exposures lead to only one outcome. Should an association be highly specific, this provides support for causality. However, since many exposures—eg, cigarette smoke—lead to numerous outcomes, lack of specificity does not argue against causation. Biological plausibility is another weak criterion, limited by our lack of knowledge. 300 years ago, clinicians would have rejected the suggestion that citrus fruits could prevent scurvy or that mosquitoes were linked with blackwater fever. Ancillary biological evidence that is coherent with the association might be helpful. For example, the effect of cigarette

Panel 2: Criteria for judgment of causal associations^{17,42,43}

Temporal sequence

Did exposure precede outcome?

Strength of association

How strong is the effect, measured as relative risk or odds ratio?

Consistency of association

Has effect been seen by others?

Biological gradient (dose-response relation)

Does increased exposure result in more of the outcome?

Specificity of association

Does exposure lead only to outcome?

Biological plausibility

Does the association make sense?

Coherence with existing knowledge

Is the association consistent with available evidence?

Experimental evidence

Has a randomised controlled trial been done?

Analogy

Is the association similar to others?

smoke on the bronchial epithelium of animals is coherent with an increased risk of cancer in human beings. Finally, experimental evidence is seldom available, and reasoning by analogy has sometimes caused harm. Since thalidomide can cause birth defects, for instance, some lawyers (successfully) argued by analogy that Bendectin (an antiemetic widely used for nausea and vomiting in pregnancy) could also cause birth defects, despite evidence to the contrary.⁴⁶

Conclusion

Studies need to have both internal and external validity: the results should be both correct and capable of extrapolation to the population. A simple checklist for bias (selection, information, and confounding) then chance can help readers decipher research reports. When a statistical association appears in research, guidelines for judgment of associations can help a reader decide whether the association is bogus, indirect, or real.

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of this material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

- Grimes DA. Technology follies: the uncritical acceptance of medical innovation. *JAMA* 1993; **269**: 3030–33.
- Last JM, ed. A dictionary of epidemiology, 2nd edn. New York: Oxford University Press, 1988.
- Ahlbom A, Norell S. Introduction to modern epidemiology, 2nd edn. Chestnut Hill, Massachusetts: Epidemiology Resources, 1990.
- Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983; **309**: 1358–61.
- Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders Company, 1985.
- Anon. The National Diet-Heart Study Final Report. *Circulation* 1968; **37**: 11–428.
- Moynour CM, Lovato LC, Thompson IM Jr, et al. Profile of men randomized to the prostate cancer prevention trial: baseline health-related quality of life, urinary and sexual functioning, and health behaviors. *J Clin Oncol* 2000; **18**: 1942–53.
- Halbert JA, Silagy CA, Finucane P, Withers RT, Hamdorf PA. Recruitment of older adults for a randomized, controlled trial of exercise advice in a general practice setting. *J Am Geriatr Soc* 1999; **47**: 477–81.
- Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998; **352**: 609–13.
- Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273**: 408–12.
- Rothman KJ. Modern epidemiology. Boston: Little, Brown and Company, 1986.
- Grimes DA. The case for confidence intervals. *Obstet Gynecol* 1992; **80**: 865–66.
- Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 “negative” trials. *N Engl J Med* 1978; **299**: 690–94.
- Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994; **272**: 122–24.
- Sackett DL. Bias in analytic research. *J Chronic Dis* 1979; **32**: 51–63.
- Wingo PA, Higgins JE, Rubin GL, Zahniser SC, eds. An epidemiologic approach to reproductive health. Geneva: WHO, 1994.
- Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Company, 1987.
- Burkman RT. Association between intrauterine device and pelvic inflammatory disease. *Obstet Gynecol* 1981; **57**: 269–76.
- Kronmal RA, Whitney CW, Mumford SD. The intrauterine device and pelvic inflammatory disease: the Women’s Health Study reanalyzed. *J Clin Epidemiol* 1991; **44**: 109–22.
- Feinstein AR, Horwitz RI. Oestrogen treatment and endometrial carcinoma. *BMJ* 1977; **2**: 766–67.
- Seltzer CC, Bosse R, Garvey AJ. Mail survey response by smoking status. *Am J Epidemiol* 1974; **100**: 453–57.
- Schull WJ, Cobb S. The intrafamilial transmission of rheumatoid arthritis: 3, the lack of support for a genetic hypothesis. *J Chronic Dis* 1969; **22**: 217–22.
- Bartholomew LL, Grimes DA. The alleged association between induced abortion and risk of breast cancer: biology or bias? *Obstet Gynecol Surv* 1998; **53**: 708–14.
- Lindefors-Harris BM, Eklund G, Adami HO, Meirik O. Response bias in a case-control study: analysis utilizing comparative data concerning legal abortions from two independent Swedish studies. *Am J Epidemiol* 1991; **134**: 1003–08.
- Harris BM, Eklund G, Meirik O, Rutqvist LE, Wiklund K. Risk of cancer of the breast after legal abortion during first trimester: a Swedish register study. *BMJ* 1989; **299**: 1430–32.
- Melbye M, Wohlfahrt J, Olsen JH, et al. Induced abortion and the risk of breast cancer. *N Engl J Med* 1997; **336**: 81–85.
- Abramson JH. Making sense of data. New York: Oxford University Press, 1988.
- Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman RB, eds. Designing clinical research: an epidemiologic approach, 2nd edn. Baltimore: Lippincott Williams and Wilkins, 2001.
- Ory HW. Association between oral contraceptives and myocardial infarction: a review. *JAMA* 1977; **237**: 2619–22.
- Schwingsl PJ, Ory HW, Visness CM. Estimates of the risk of cardiovascular death attributable to low-dose oral contraceptives in the United States. *Am J Obstet Gynecol* 1999; **180**: 241–49.
- Jain AK. Cigarette smoking, use of oral contraceptives, and myocardial infarction. *Am J Obstet Gynecol* 1976; **126**: 301–07.
- Lee NC, Rubin GL, Ory HW, Burkman RT. Type of intrauterine device and the risk of pelvic inflammatory disease. *Obstet Gynecol* 1983; **62**: 1–6.
- Lee NC, Rubin GL, Borucki R. The intrauterine device and pelvic inflammatory disease revisited: new results from the Women’s Health Study. *Obstet Gynecol* 1988; **72**: 1–6.
- Lee NC, Rubin GL, Grimes DA. Measures of sexual behavior and the risk of pelvic inflammatory disease. *Obstet Gynecol* 1991; **77**: 425–30.
- Schlesselman JJ. Cancer of the breast and reproductive tract in relation to use of oral contraceptives. *Contraception* 1989; **40**: 1–38.
- Lacey JV Jr, Brinton LA, Abbas FM, et al. Oral contraceptives as risk factors for cervical adenocarcinomas and squamous cell carcinomas. *Cancer Epidemiol Biomarkers Prev* 1999; **8**: 1079–85.
- Kjellberg L, Hallmans G, Ahren AM, et al. Smoking, diet, pregnancy and oral contraceptive use as risk factors for cervical intra-epithelial neoplasia in relation to human papillomavirus infection. *Br J Cancer* 2000; **82**: 1332–38.
- Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959; **22**: 719–48.
- Lang TA, Secic M. How to report statistics in medicine. Philadelphia: American College of Physicians, 1997.
- Rothman KJ. A show of confidence. *N Engl J Med* 1978; **299**: 1362–63.
- Sterne JA, Smith GD. Sifting the evidence: what’s wrong with significance tests? *BMJ* 2001; **322**: 226–31.
- Hill AB. The environment and disease association or causation. *Proc R Soc Med* 1965; **58**: 295–300.
- Streiner DL, Norman GR, Munroe Blum H. PDQ epidemiology. Toronto: BC Decker, 1989.
- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine, 2nd edn. Boston: Little, Brown and Company, 1991.
- Grimes DA, Economy KE. Primary prevention of gynecologic cancers. *Am J Obstet Gynecol* 1995; **172**: 227–35.
- McKeigue PM, Lamm SH, Linn S, Kutcher JS. Bendectin and birth defects: 1, a meta-analysis of the epidemiologic studies. *Teratology* 1994; **50**: 27–37.

Cohort studies: marching towards outcomes

David A Grimes, Kenneth F Schulz

A cohort study tracks two or more groups forward from exposure to outcome. This type of study can be done by going ahead in time from the present (prospective cohort study) or, alternatively, by going back in time to comprise the cohorts and following them up to the present (retrospective cohort study). A cohort study is the best way to identify incidence and natural history of a disease, and can be used to examine multiple outcomes after a single exposure. However, this type of study is less useful for examination of rare events or those that take a long time to develop. A cohort study should provide specific definitions of exposures and outcomes: determination of both should be as objective as possible. The control group (unexposed) should be similar in all important respects to the exposed, with the exception of not having the exposure. Observational studies, however, rarely achieve such a degree of similarity, so investigators need to measure and control for confounding factors. Reduction of loss to follow-up over time is a challenge, since differential losses to follow-up introduce bias. Variations on the cohort theme include the before-after study and nested case-control study (within a cohort study). Strengths of a cohort study include the ability to calculate incidence rates, relative risks, and 95% CIs. This format is the preferred way of presenting study results, rather than with *p* values.

The term cohort has military, not medical, roots. A cohort was a 300–600-man unit in the Roman army; ten cohorts formed a legion (figure 1). The etymology of the term provides a useful mnemonic: a cohort study consists of bands or groups of persons marching forward in time from an exposure to one or more outcomes.

This analogy might be helpful, since cohort studies have a bevy of confusing synonyms: incidence, longitudinal, forward-looking, follow-up, concurrent, and prospective study.^{1,2} Although the terminology can seem daunting, the cohort study is easy for clinicians to understand, since it flows in a logical direction (unlike the case-control study). Here, we explain the terminology, describe the strengths and weaknesses of cohort studies, consider several logistical concerns, mention two permutations of cohort studies, and summarise their analysis.

Data collection: forwards and backwards

A cohort study follows-up two or more groups from exposure to outcome. In its simplest form, a cohort study compares the experience of a group exposed to some factor with another group not exposed to the factor. If the former group has a higher or lower frequency of an outcome than the unexposed, then an association between exposure and outcome is evident.

The defining characteristic of all cohort studies is that they track people forward in time from exposure to outcome. Researchers doing this kind of study must, therefore, go forward in time from the present or go back in time to choose their cohorts (figure 2). Either way, a cohort study moves in the same direction, although gathering data might not. For example, an investigator who wants to study the epidemic of multiple births stemming from assisted reproductive technologies³ could begin a cohort study now. Women exposed to these technologies and a similar group who conceived naturally



Figure 1: An early cohort in search of favourable outcomes

could be tracked forward through their pregnancies to monitor the frequency of multiple births (a concurrent cohort study). Alternatively, the investigator might use existing medical records and go back in time several years to identify women exposed and not exposed to these technologies. He would then track them forward through records to note the birth outcomes. Again, the study moves from exposure to outcome, though the data collection occurred after the fact.

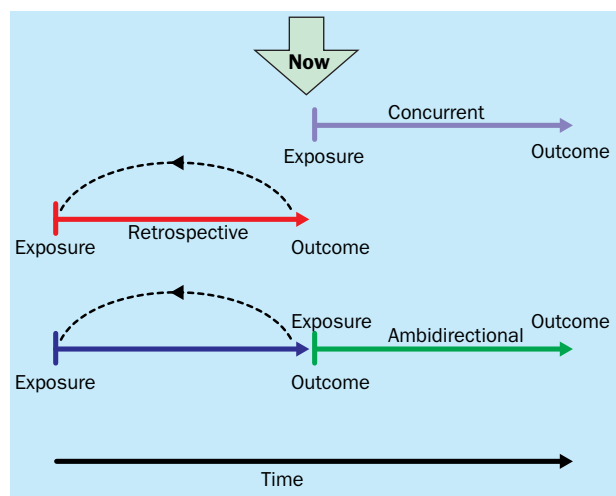


Figure 2: Schematic diagram of concurrent, retrospective, and ambidirectional cohort studies

Lancet 2002; **359**: 341–45

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (D A Grimes MD, K F Schulz PhD)

Correspondence to: Dr David A Grimes (e-mail: dgrimes@fhi.org)

Yet a third variation exists: ambidirectional.⁴ As the name implies, data collection goes in both directions. This approach can be useful for exposures that have both short-term and long-term outcomes. In this hypothetical example, assisted reproductive technologies might be associated with multiple births and with ovarian cancer in later life.⁵ The investigator might, therefore, look back through records for multiple births and also start to follow-up these women into the future for ovarian cancer occurrence.

Advantages of cohort studies

Cohort studies have many appealing features. They are the best way to ascertain both the incidence and natural history of a disorder.⁶ The temporal sequence between putative cause and outcome is usually clear: the exposed and unexposed can often be seen to be free of the outcome at the outset. By contrast, this chicken-egg question often frustrates cross-sectional and case-control studies. For example, in a case-control study, patients with chronic widespread pain were more likely to have mental illness than controls.⁷ Do mood and anxiety disorders increase this risk, or do patients with chronic pain develop mood and anxiety disorders as a result of their disorder?

Cohort studies are useful in investigation of multiple outcomes that might arise after a single exposure. A prototype would be cigarette smoking (the exposure) and stroke, emphysema, oral cancer, and heart disease (the outcomes). Although assessment of many outcomes is often cited as a positive attribute of cohort studies, this feature can be abused. For example, testing the associations between exposure and many outcomes, but only reporting the significant ones, represents misleading science. Investigators should preferably have planned primary and secondary associations to examine (sometimes called hypothesis confirmation). Although investigators can look at other outcomes (hypothesis generation), they should report the findings of all examinations, not just significant ones, so that readers can correctly interpret the results.

The cohort design is also useful in the study of rare exposures: a researcher can often recruit people with uncommon exposures—eg, to ionising radiation or chemicals—in the workplace. A hospital or factory might provide a large number of individuals with the exposure of interest, which would be rare in the general population. Since the investigator does not assign exposure, no ethical concerns arise.

Cohort studies also reduce the risk of survivor bias.⁶ Diseases that are rapidly fatal are difficult to study because of this factor. For example, a hospital-based case-control study of the link between snow-shovelling and myocardial infarction would miss all those who died in the driveway. A cohort study would be a less biased (but more cumbersome) approach: compare rates of myocardial infarction among those who shovel and those who do not shovel. Finally, cohort studies allow calculation of incidence rates, relative risks, and confidence intervals.² Other outcome measures in cohort studies include life-table rates, survival curves, and hazard ratios (panel 1).^{8–10} By contrast, case-control studies cannot provide incidence rates; at best, odds ratios approximate relative risks only when the outcome is uncommon.

Disadvantages of cohort studies

Cohort studies have important limitations too. Selection bias is built into cohort studies. For example, in a cohort study investigating effects of jogging on cardiovascular

Panel 1: Reporting time-to-event in cohort studies

Survival analysis

Survival analysis is useful when lengths of follow-up vary substantially or when participants enter a study at different times.⁸ The Kaplan-Meier method provides a more sophisticated expression of the risk of the outcome over time than does a simple dichotomous outcome.⁹ It can determine the probability (P) of the outcome at any point in time; this result is graphed as a step function (which jumps at every event). A complementary, mirror-image graph portrays the likelihood of avoiding the outcome (1–P) as a function of time (Kaplan-Meier survival curve). The log-rank test compares survival curves of different groups.¹⁰

Proportional hazard model

Another approach to different lengths of follow-up is the Cox proportional hazard model. It is a multivariate technique that has time-to-event (such as illness) as the dependent variable. By contrast, multiple logistic regression has "yes-no" as the dependent variable.⁸ Coefficients from this model can be used to calculate the risk ratio (hazard ratio) of the outcome, after controlling for other covariates in the equation. The hazard ratio (with 95% CIs) is interpreted in the same way as a relative risk for dichotomous outcomes.¹⁰

disease, those who choose to jog probably differ in other important ways (such as diet and smoking) from those who do not exercise.¹¹ In theory, both groups should be the same in all important respects, except for the exposure of interest (jogging), but this seldom occurs. The cohort design is not optimum for rare diseases—eg, scleroderma—or those that take a long time to develop—eg, cancer. However, several large (and thus expensive) cohort studies have made landmark contributions to our knowledge of uncommon diseases. Examples include the Royal College of General Practitioners' Oral Contraceptive Study,¹² the Framingham Heart Study,¹³ the Nurses Health Study,¹⁴ and the British Physicians' Study.¹⁵

Loss to follow-up can be a difficulty, even at 1 month, and particularly so with longitudinal studies that continue for decades. Differential losses to follow-up between those exposed and unexposed can bias results. Over time, the exposure status of study participants can change. For example, a proportion of women who use oral contraceptives will switch to an intrauterine device, and vice versa.¹⁶ Partitioning might be needed to avoid a blurring of exposure, sometimes termed contamination.

What to look for in cohort studies

Who is at risk?

All participants (both exposed and unexposed) in a cohort study must be at risk of developing the outcome.⁶ For example, since women who have had a tubal sterilisation operation have almost no risk of salpingitis,¹⁷ they should not be included in cohort studies of pelvic inflammatory disease.

Who is exposed?

Cohort studies need a clear, unambiguous definition of the exposure at the outset. This definition sometimes involves quantifying the exposure by degree, rather than just yes or no. For example, the minimum exposure might have to be 14 cigarettes per day or less,¹⁸ or 3–6 months of oral contraceptives.¹⁹ Definition of exposure levels in this way can result in more than two groups—eg, non-smokers, light smokers, and heavy smokers.¹⁸

Who is an appropriate control?

The key notion is that controls (the unexposed) should be similar to the exposed in all important respects, except for the lack of exposure. If so, the unexposed group will reveal the background rate of the outcome in the community.

The unexposed group can come from either internal (persons from the same time and place, such as a hospital ward) or external sources. Internal comparisons are most desirable. In a particular population, individuals segregate by themselves (or through medical interventions) into exposure status—eg, cigarette smoking, occupation, contraception. For example, in a cohort study, 138 patients with HIV-1-associated Kaposi's sarcoma were divided into two groups: those with oral and those with cutaneous lesions. The presence of oral lesions (the exposure) had a poorer prognosis, with a median survival (the outcome) one-third that of the other group.²⁰

If satisfactory internal controls are not available, researchers look elsewhere (sometimes termed a double-cohort study).⁶ In a trial of an occupational exposure, finding an adequate number of employees in the factory without the exposure might be difficult. Hence, one might choose workers in a similar factory in the same community. This choice assumes that workers in the other factory have the same baseline risk of the outcome in question, which might not be the case. Even less desirable is use of population norms; disease-specific mortality rates are an example. A researcher might compare lung-cancer death rates among workers in the factory with rates of persons of the same age and sex in the population. Bias inevitably creeps into such comparisons because of the healthy worker effect: those who work are healthier, in general, than those who do not (or cannot) work.^{4,9} Additionally, work reaps economic benefits which might further bias comparisons.

Have outcomes been assessed equally?

Outcomes must be defined in advance; they should be clear, specific, and measurable. Identification of outcomes should be comparable in every way for the exposed and unexposed to avoid information bias. Failure to define objective outcomes leads to uninterpretable results. This challenge relates not only to subjective syndromes such as Gulf War,²¹ chronic fatigue,^{22,23} and premenstrual,²⁴ but also to more mundane health problems such as endometritis. Just how tender must a uterus be? Keeping those who judge outcomes unaware of the exposure status of participants (blinding) in a cohort study is important for subjective outcomes, such as tenderness or erythema. By contrast, with objective outcome measures, such as fever or death, blinding the exposure status is less important.

Outcome information can come from many sources. For mortality studies, the death certificate is often used. Although convenient, the validity of the clinical information is highly variable. For non-fatal outcomes, sources include hospital charts, insurance records, laboratory records, disease registries, hospital discharge logs, and physical examination and measurement of participants. Optimally, the person who judges outcomes should be unaware of the exposure. When diagnoses vary in their confidence, assignment of levels of assurance might be helpful, such as definite, probable, and suspect.⁹

Tracking participants over time*Have losses been minimised?*

Although loss of participants damages the power and precision of a study, differential loss to follow-up is more sinister. Bail-outs are not random events. If the likelihood

of bailing out is related both to exposure and outcome, then bias can result.²⁵ For example, some participants given a new antibiotic might have such poor outcomes that they are unable to complete questionnaires or to return for examination.²⁶ Their disappearance from the cohort would make the new antibiotic look better than it is.

The best way of dealing with loss to follow-up is to avoid it. For example, restrict participation to only those judged likely to complete the study. Additionally, several safeguards are customary. Obtaining the names of several family members or friends who do not live with the respondent is often helpful at the start of such studies. The participant's family doctor might also be helpful. Should the respondent move, these contacts would probably know their new address. Motor vehicle registration records can be useful too. Furthermore, national vital statistics registries, such as the National Death Index in the USA, facilitate follow-up. Participants can be offered financial compensation for their time lost from work as a result of the study. Diligent tracking of participants is hard work, and might require hiring personnel for this task alone.

Reporting cohort studies

Many researchers who do cohort studies report their findings in an unsatisfactory way (panel 2).²⁷ An investigator's first challenge is to convince the editor (then readers) that the exposed and unexposed groups were indeed similar in all important respects, except for the exposure. The first table in reports of cohort studies customarily provides demographic and other prognostic factors for both groups with hypothesis testing (p values) to show the likelihood that observed differences could be due to chance.

For dichotomous outcome measures, such as sick or well, the investigator should provide raw data sufficient for the reader to confirm the results. For cumulative incidence, the investigator should calculate the proportion who developed the outcome during the specified study interval. For incidence rates, the value is expressed per unit of time.⁴ Then, relative risks and confidence intervals should be provided. Use of p values should not replace interval estimation (relative risks with confidence

Panel 2: Features to look for in a cohort study**How much selection bias was present?**

- 1 Were only people at risk of the outcome included?
- 1 Was the exposure clear, specific, and measurable?
- 1 Were the exposed and unexposed groups similar in all important respects except for the exposure?

What steps were taken to minimise information bias?

- 1 Was the outcome clear, specific, and measurable?
- 1 Was the outcome identified in the same way for both groups?
- 1 Was determination of outcome made by an observer blinded as to treatment?

How complete was the follow-up of both groups?

- 1 What efforts were made to limit loss to follow-up?
- 1 Was loss to follow-up similar in both groups?

Were potential confounding factors sought and controlled for in the analysis?

- 1 Did the investigators anticipate and gather information on potential confounding factors?
- 1 What method(s) were used to assess and control for confounding?

intervals)²⁸ and should only be used as supplemental information.²⁵

Like other observational studies, cohort studies have built-in bias. Investigators should identify potential biases in their data and show how these might have affected results. Whenever possible, confounding should be controlled for in the analysis. These techniques are discussed in an earlier essay in this series.²⁹

Variations on the cohort theme

Before-after studies

Before-after studies (time series) have important limitations. Here, an investigator takes a measurement, exposes participants to an intervention (often a drug), repeats the measurements, then compares them. First, regression to the mean is often ignored. If admission to the cohort includes extreme measurements,³⁰ such as high laboratory values, then lower mean values will arise at follow-up, irrespective of treatment.³¹ Second, secular trends, such as seasonal changes in the frequency of pneumonia, can affect results. Third, washout periods are often needed to avoid a carryover effect of drugs given during the initial observation period.⁶

Nested case-control studies

Cohort studies sometimes spawn other studies. One of the most frequent is the nested case-control study.^{6,9,25} Why would an investigator carve out a case-control study in the midst of a cohort study? The answer often involves body fluids and a freezer. Some exposure or predictor variables are simply too expensive to determine on everyone in a study. A sophisticated blood test is the prototype. A clever way to skirt this financial obstacle is to do a cohort study that will yield a sufficient number of cases. All participants entering the cohort study have a tube of blood drawn at enrolment; serum is frozen until the study's conclusion. All those in the cohort study who develop the outcome of interest now become the cases for the nested study. The investigator then chooses a random sample of all participants who did not develop the outcome (controls). Next, the blood test is done on serum from only the cases and controls, not the whole group of exposed and unexposed. In this way, the laboratory cost is minimised while assuring that the exposure—eg, a positive laboratory test—was present before development of the outcome. Controls are generally matched to cases by important characteristics, such as age and sex.⁹

A nested case-control study, for example, examined the potential relation between body concentrations of organochlorines and non-Hodgkin's lymphoma. The blood samples were obtained on entry to a large cohort study started in Maryland, USA, in 1974. Blood samples were eventually analysed for only 74 individuals with lymphoma and 147 controls.³² Thus, instead of measuring organochlorine concentrations of the entire cohort of 25 802, the investigators incurred this laboratory expense for less than 1% of the cohort. In view of the availability of banked blood specimens around the world, this type of research design is likely to become popular. However, nested case-control studies might be useful for other studies that do not require blood tests but in which determination of the exposure is expensive or difficult⁹—eg, measurement of nerve conduction³³ or job stressors.³⁴

Conclusion

Cohort studies are common in medical research. Like other research designs, they entail important trade-offs. Readers should make sure that investigators provide clear,

specific, and measurable definitions of exposures and outcomes. The unexposed group should resemble the exposed group in all important respects, and determination of outcomes should be objective and, whenever possible, blinded. Results for dichotomous outcomes should be provided as rates, relative risks, and confidence intervals, which offer more information than do p values. Reports of cohort studies should identify and describe the potential effect of biases. Importantly, investigators should measure and control for potential confounding.

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of this material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

- 1 Last JM, ed. A dictionary of epidemiology, 2nd edn. New York: Oxford University Press, 1988.
- 2 Lilienfeld AM, Lilienfeld DE. Foundations of epidemiology, 2nd edn. New York: Oxford University Press, 1980.
- 3 Anon. Contribution of assisted reproductive technology and ovulation-inducing drugs to triplet and higher-order multiple births—United States, 1980–1997. *MMWR Morb Mortal Wkly Rep* 2000; **49**: S35–38.
- 4 Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Company, 1987.
- 5 Whittemore AS, Harris R, Itnyre J. Characteristics relating to ovarian cancer risk: collaborative analysis of 12 US case-control studies, II: invasive epithelial ovarian cancers in white women. Collaborative Ovarian Cancer Group. *Am J Epidemiol* 1992; **136**: 1184–203.
- 6 Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman RB, eds. Designing clinical research: an epidemiologic approach, 2nd edn. Baltimore: Lippincott Williams and Wilkins, 2001.
- 7 Benjamin S, Morris S, McBeth J, Macfarlane GJ, Silman AJ. The association between chronic widespread pain and mental disorder: a population-based study. *Arthritis Rheum* 2000; **43**: 561–67.
- 8 Wassertheil-Smoller S. Biostatistics and epidemiology: a primer for health professionals, 2nd edn. New York: Springer-Verlag, 1995.
- 9 Kelsey JL, Whittemore AS, Evans AS, Thompson WD. Methods in observational epidemiology, 2nd edn. New York: Oxford University Press, 1996.
- 10 Lang TA, Secic M. How to report statistics in medicine. Philadelphia: American College of Physicians, 1997.
- 11 Sackett DL. Bias in analytic research. *J Chron Dis* 1979; **32**: 51–63.
- 12 Beral V, Hermon C, Kay C, Hannaford P, Darby S, Reeves G. Mortality associated with oral contraceptive use: 25 year follow up of cohort of 46 000 women from Royal College of General Practitioners' oral contraception study. *BMJ* 1999; **318**: 96–100.
- 13 Seman LJ, DeLuca C, Jenner JL, et al. Lipoprotein(a)-cholesterol and coronary heart disease in the Framingham Heart Study. *Clin Chem* 1999; **45**: 1039–46.
- 14 Colditz GA, Rosner BA, Speizer FE. Risk factors for breast cancer according to family history of breast cancer. *J Natl Cancer Inst* 1996; **88**: 365–71.
- 15 Doll R, Peto R, Boreham J, Sutherland I. Smoking and dementia in male British doctors: prospective study. *BMJ* 2000; **320**: 1097–102.
- 16 Hannaford PC, Kay CR. The risk of serious illness among oral contraceptive users: evidence from the RCGP's oral contraceptive study. *Br J Gen Pract* 1998; **48**: 1657–62.
- 17 Levgur M, Duvivier R. Pelvic inflammatory disease after tubal sterilization: a review. *Obstet Gynecol Surv* 2000; **55**: 41–50.
- 18 Croft P, Hannaford P. Risk factors for acute myocardial infarction in women. *BMJ* 1989; **298**: 674.
- 19 Anon. The reduction in risk of ovarian cancer associated with oral-contraceptive use. The Cancer and Steroid Hormone Study of the Centers for Disease Control and the National Institute of Child Health and Human Development. *N Engl J Med* 1987; **316**: 650–55.
- 20 Rohrmus B, Thoma-Greber EM, Bogner JR, Rocken M. Outlook in oral and cutaneous Kaposi's sarcoma. *Lancet* 2000; **356**: 2160.
- 21 Fukuda K, Nisenbaum R, Stewart G, et al. Chronic multisymptom illness affecting Air Force veterans of the Gulf War. *JAMA* 1998; **280**: 981–98.
- 22 Schluenderberg A, Straus SE, Peterson P, et al. NIH conference. Chronic fatigue syndrome research: definition and medical outcome assessment. *Ann Intern Med* 1992; **117**: 325–31.
- 23 Armon C, Kurland LT. Chronic fatigue syndrome: issues in the diagnosis and estimation of incidence. *Rev Infect Dis* 1991; **13** (suppl 1): S68–72.

- 24 Haskett RF, Abplanalp JM. Premenstrual tension syndrome: diagnostic criteria and selection of research subjects. *Psychiatr Res* 1983; **9**: 125–38.
- 25 Rothman KJ. *Modern epidemiology*. Boston: Little, Brown and Company, 1986.
- 26 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*, 2nd edn. Boston: Little, Brown and Company, 1991.
- 27 Bracken MB. Reporting observational studies. *Br J Obstet Gynaecol* 1989; **96**: 383–88.
- 28 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ* 1986; **292**: 746–50.
- 29 Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet* 2002; **359**: 248–52.
- 30 Ewan PW, Clark AT. Long-term prospective observational study of patients with peanut and nut allergy after participation in a management plan. *Lancet* 2001; **357**: 111–15.
- 31 Bland JM, Altman DG. Regression towards the mean. *BMJ* 1994; **308**: 1499.
- 32 Rothman N, Cantor KP, Blair A, et al. A nested case-control study of non-Hodgkin lymphoma and serum organochlorine residues. *Lancet* 1997; **350**: 240–44.
- 33 Atterbury MR, Limke JC, Lemasters GK, et al. Nested case-control study of hand and wrist work-related musculoskeletal disorders in carpenters. *Am J Ind Med* 1996; **30**: 695–701.
- 34 Kawakami N, Araki S, Kawashima M. Effects of job stress on occurrence of major depression in Japanese industry: a case-control study nested in a cohort study. *J Occup Med* 1990; **32**: 722–25.

Uses of error

The error cascade

Neil Gittoes

Having just been appointed consultant physician, I found myself reflecting on my career and realising that I could at last stand alone and that finally the buck stops with me. There was a time that I wished it didn't. I was a medical senior house officer when I saw an elderly man who described a subacute onset of breathlessness and a dry cough. He had trouble speaking and was using his accessory muscles. He had initially received standard nebulised treatment for exacerbated chronic obstructive pulmonary disease, although his chest radiograph showed a small pneumothorax on the left. After conferring with senior colleagues, I inserted a chest drain on the left, and verified its position with a second radiograph. In the middle of the night the house officer saw the patient with worsening shortness of breath and surgical emphysema. He pushed the tube in further, but an hour later the arrest team were called because the patient had developed extreme respiratory distress and had become cyanosed. They thought that he had developed a contralateral

pneumothorax and proceeded to insert a chest drain on the right. Arriving on the ward the following morning, I was horrified to find my patient with bilateral chest drains and surgical emphysema from head to scrotum. However, at least he was alive. Chest radiographs and computed tomography showed bilateral pneumothoraces with both drains embedded deeply within the lung parenchyma, just short of the mediastinum on the right, and abutting the left ventricle on the left. I inserted bilateral anterior drains and cautiously removed the lateral ones. After a few days the right-sided pneumothorax resolved, although the left side needed surgical correction. He was finally discharged, and on reviewing the radiographs it was apparent that I had inserted the original drain where there was a small area of pleural adhesion. The two pleural surfaces remained contiguous, and the drain entered the lung parenchyma. The subsequent errors of management turned the situation rapidly into a life-threatening predicament.

Queen Elizabeth Hospital, Edgbaston, Birmingham, B15 2 TH UK (N Gittoes MRCP)

Case-control studies: research in reverse

Kenneth F Schulz, David A Grimes

Epidemiologists benefit greatly from having case-control study designs in their research armamentarium. Case-control studies can yield important scientific findings with relatively little time, money, and effort compared with other study designs. This seemingly quick road to research results entices many newly trained epidemiologists. Indeed, investigators implement case-control studies more frequently than any other analytical epidemiological study. Unfortunately, case-control designs also tend to be more susceptible to biases than other comparative studies. Although easier to do, they are also easier to do wrong. Five main notions guide investigators who do, or readers who assess, case-control studies. First, investigators must explicitly define the criteria for diagnosis of a case and any eligibility criteria used for selection. Second, controls should come from the same population as the cases, and their selection should be independent of the exposures of interest. Third, investigators should blind the data gatherers to the case or control status of participants or, if impossible, at least blind them to the main hypothesis of the study. Fourth, data gatherers need to be thoroughly trained to elicit exposure in a similar manner from cases and controls; they should use memory aids to facilitate and balance recall between cases and controls. Finally, investigators should address confounding in case-control studies, either in the design stage or with analytical techniques. Devotion of meticulous attention to these points enhances the validity of the results and bolsters the reader's confidence in the findings.

Case-control studies contribute greatly to the research toolbox of an epidemiologist. They embody the strengths and weaknesses of observational epidemiology. Moreover, epidemiologists use them to study a huge variety of associations. To show this variety, we searched PubMed for topics investigated with case-control studies (panel 1).^{1–20} We identified diverse diseases and exposures, with outcomes ranging from earthquake deaths to racehorse injuries, and exposures ranging from pickled vegetables to pig farming.

The strength of case-control studies can be appreciated in early research done by investigators hoping to understand the cause of AIDS. Case-control studies identified risk groups—eg, homosexual men, intravenous drug users, and blood-transfusion recipients—and risk factors—eg, multiple sex partners, receptive anal intercourse in homosexual men, and not using condoms—for AIDS. Based on such studies, blood banks restricted high-risk individuals from donating blood, and educational programmes began to promote safer behaviours. As a result of these precautions, the speed of transmission of HIV-1 was greatly reduced, even before the virus had been identified.

By comparison with other study types, case-control studies can yield important findings in a relatively short time, and with relatively little money and effort. This apparently quick road to research results entices many newly trained epidemiologists. However, case-control studies tend to be more susceptible to biases than other analytical, epidemiological designs.²¹ A notable friend of ours (David L Sackett, personal communication) told us that he would trust only six people in the world to do a proper case-control study. And, in his book, Rothman comments that: “because it need not be extremely

expensive nor time-consuming to conduct a case-control study, many studies have been conducted by would-be investigators who lack even a rudimentary appreciation for epidemiologic principles. Occasionally such haphazard research can produce fruitful or even extremely important results, but often the results are wrong because basic research principles have been violated.”²²

Panel 1: Examples of topics investigated with case-control studies

Exposure	Outcome
Cat ownership in childhood	Schizophrenia, schizoaffective disorder, or bipolar disorder ¹
Body-mass index	Pancreatic cancer ²
Physical disability	Earthquake mortality ³
Hiatus hernia	Reflux oesophagitis ⁴
Hair dyes	Connective tissue disorders ⁵
History of shingles	Systemic lupus erythematosus ⁶
Pig farming	Nipah virus infection ⁷
Ghee (clarified butter) applied to umbilical cord stump	Neonatal tetanus ⁸
Pickled vegetable consumption	Oesophageal cancer ⁹
Turf running surface	Musculoskeletal injury in thoroughbred racehorses ¹⁰
Digital rectal exam	Metastatic prostate cancer ¹¹
Statins for lipid lowering	Dementia ¹²
Paracetamol use	Ovarian cancer ¹³
Phyto-oestrogens	Breast cancer ¹⁴
Overhead mirror at intersections	Forklift collision injuries ¹⁵
Male condom use	Genital warts ¹⁶
Physical activity	Ovarian cancer ¹⁷
Sigmoidoscopy screening	Colon cancer ¹⁸
Large doses of folate and iron in pregnancy	Microcephaly ¹⁹
Influenza vaccination	Recurrent myocardial infarction ²⁰

Lancet 2002; **359**: 431–34

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (K F Schulz PhD, D A Grimes MD)

Correspondence to: Dr Kenneth F Schulz (e-mail: KSchulz@fhi.org)

Basic case-control study design

Case-control designs might seem easy to understand, but many clinicians stumble over them. Because this type of study runs backwards by comparison with most other studies, it often confuses researchers and readers alike. In cohort studies, for example, study groups are defined by exposure. In case-control studies, however, study groups are defined by outcome (figure). To study the association between smoking and lung cancer, therefore, people with lung cancer are enrolled to form the case group, and people without lung cancer are identified as controls.

Researchers then look back in time to ascertain each person's exposure status (smoking history), hence the retrospective nature of this study design. Investigators compare the frequency of smoking exposure in the case group with that in the control group, and calculate a measure of association.²¹⁻²³

Unlike cohort studies, case-control studies cannot yield incidence rates.²⁴ Instead, they provide an odds ratio, derived from the proportion of individuals exposed in each of the case and control groups. When the incidence rate of a particular outcome in the population of interest is low (usually under 5% in both the exposed and unexposed suffices)²¹ the odds ratio from a case-control study is a good estimate of relative risk.^{21,23}

Advantages and disadvantages

Epidemiologists often tout case-control studies as the most efficient design in terms of time, money, and effort. This recommendation makes sense when the incidence rate of an outcome is low, since in a cohort design the researchers would have to follow up many individuals to identify one with the outcome. Case-control studies are also efficient in the investigation of diseases that have a long latency period—eg, cancer—in which instance a cohort study would involve many years of follow-up before the outcome became evident.

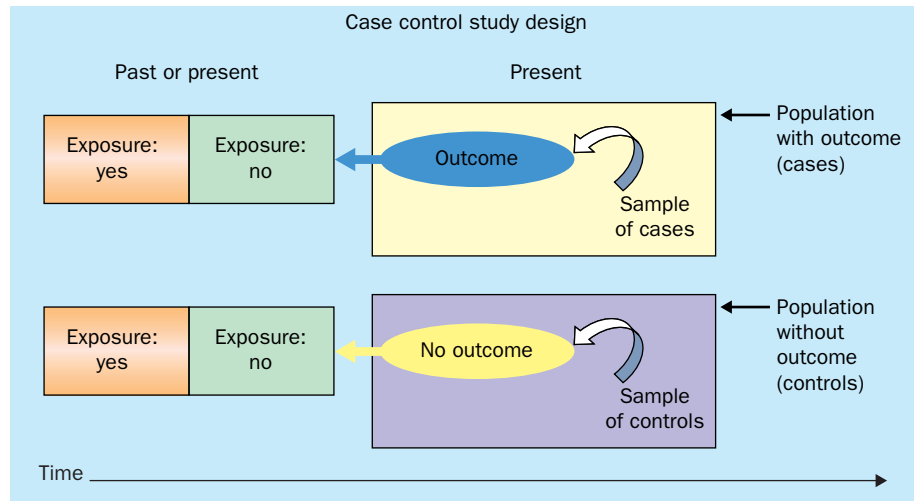
However, cohort studies can be more efficient than case-control studies. If the frequency of exposure is low, for example, case-control studies quickly become inefficient. Researchers would have to examine many cases and controls to find one who had been exposed. For instance, a case-control study of oral contraceptive use and transmission of HIV-1 would be impractical in parts of Africa because of the rarity of use of oral contraceptives. As a rule of thumb, cohort designs are more efficient in settings in which the incidence of outcome is higher than the prevalence of exposure.

Finally, many methodological issues affect the validity of the results of case-control studies, and two factors—ie, choosing a control group and obtaining exposure history—can greatly affect a study's vulnerability to bias.

Selection of case and control groups

Case group

All the cases from a population could, theoretically, be included as participants in a case-control study. For practical reasons, however, only a sample is frequently studied.²² Investigators should, therefore, state how the



Schematic diagram of case-control study design

sample was selected, providing a clear definition of the outcome being studied including, for example, clinical symptoms, laboratory results, and diagnostic methods used. Furthermore, researchers should detail eligibility criteria used for selection, such as age range and location (clinic, hospital, population-based). Finally, they should gather data preferably from incident (new) rather than prevalent (both old and new) cases;²⁵ since diagnostic patterns change over time, recent diagnoses are likely to be more consistent than those obtained from different periods.

Control group

The control group provides the background proportion of exposure expected in the case group. Controls should, therefore, be free of the disease (outcome) being studied, but should be representative of those individuals who would have been selected as cases had they developed the disease. In other words, controls should represent the population at risk of becoming cases.

Selection of controls must be independent of the exposure being investigated. Artistic licence enters the study design at this point, sometimes for the better and, unfortunately, sometimes for the worse. When investigators consider potential control groups, they must anticipate all the potential biases that could arise, making this task one of the hardest in epidemiology.

Suppose investigators selected individuals with myocardial infarction from the cardiology ward of a large, city hospital as cases, but identified people without infarction from the emergency medicine ward of the same hospital as controls. Bias might result. The cardiology ward is used as a referral centre for the entire state, whereas the emergency medicine department primarily serves only the city. Unfortunately, the exposure history for patients from the city would not usually accurately reflect that of patients statewide. For example, the exposure of interest—eg, a new blood pressure drug—might not be available to patients in outlying areas of the state but be commonly prescribed in the city. In this example, therefore, either the controls should be chosen from the entire state, like the cases, or the investigators should exclude all individuals who lived outside the local community served by the emergency medicine ward. Moreover, controls should be selected independent of exposure. Assume that this new antihypertensive drug causes drowsiness and slows reaction time. Such side-effects might lead to automobile accidents, with injured

drivers entering the emergency medicine department. Thus, the investigator's control group would include an abnormally high proportion of individuals exposed to the new antihypertensive, a biased comparison with the case group.

Another hypothetical example could be a case-control study of whether non-steroidal anti-inflammatory drugs (NSAIDs) prevent colorectal cancer. The study measures previous NSAID use by patients admitted to hospital with (cases) and without (controls) colorectal cancer. If the control group came from the rheumatology service, then the study would be biased, since individuals with arthritis use NSAIDs more often than do the general population from which the cases were chosen. Such a high level of NSAID use in controls would result in a spuriously low risk (odds ratio) calculation. Alternatively, if the control group came from the gastroenterology service, where many ulcer patients had been advised by their doctors to avoid NSAIDs, then that control group might yield a low level of NSAID use and a spuriously high risk (odds ratio) calculation. In other words, if investigators do not select control groups independent of exposure, biases in either direction might result (panel 2).

An early case-control study in AIDS serves as a good example of how inappropriate controls can result in biased findings.²⁶ In this instance, the researchers compared cases of AIDS diagnosed in San Francisco, CA, USA, between 1983 and 1984 with two HIV-uninfected control groups. One control group included individuals who attended a clinic for sexually transmitted diseases (STD), and the other included people identified from the neighbourhoods of the cases. The investigators compared the risk of AIDS in individuals with more than 100 sexual partners with that in people with no to five sexual partners. The resulting odds ratios were 2.9 with STD clinic controls, but 52.0 with neighbourhood controls. The magnitude of this difference shows the potential for huge biases due to selection of improper control groups. In this study, controls from the STD clinic proved inappropriate, since their selection was not independent of exposure (more than 100 sexual partners). Acquisition of STDs is associated with number of sexual partners, thus these controls generated a highly biased odds ratio estimate.

Investigators can reduce selection bias by minimising judgment in the selection process. For example, if the case group included all affected individuals in a specified geographic region, then the control group could be chosen at random from the general population of the same area. This approach was used in a case-control study of breast cancer and oral contraceptive use.²⁷ All women aged 20–54 years, who had newly diagnosed breast cancer, and who lived in one of eight geographic areas in the USA formed the case group. Women of the same ages, selected by random digital telephone dialling, and from the same areas, formed the control group. Although this study represents an excellent example, such designs are not always feasible.

Readers of case-control studies should not accept results of studies without checking the appropriateness of the control group, as described in the methods section. If the researchers provide little insight into the choice of their control group, become sceptical. Examine whatever information the researcher has provided for indications about how well the control group represents the cases, independent of the exposure being studied.²⁵ This assessment takes time and energy, but it represents the crux of a case-control study.

Measurement of exposure information

Another difficulty in case-control studies involves the measurement of exposure information. Participants, both cases and controls, might inaccurately remember past exposures, especially those that happened a long time ago. Furthermore, cases often remember exposures to putative risk factors differently than controls. This differential recall (recall bias) causes information bias.²⁵

In the study of breast cancer and oral contraceptive use,²⁷ for example, investigators asked participants about previous exposure to oral contraceptives. Women with breast cancer might have searched their minds for what could have caused their cancers, identifying oral contraceptives as a risk because of stories in the media about the postulated relation between contraceptives and breast cancer. Thus, although some women in each group might have used a particular oral contraceptive 20 years ago, the case might remember taking it whereas the control might not. Such recall bias would generate an exaggerated relation between oral contraceptives and breast cancer. Information bias is especially pernicious because analytical techniques, irrespective of their sophistication, cannot moderate or eliminate it.

In a Swedish study,²⁸ investigators examined the potential link between induced abortion and later development of breast cancer. They gathered information about exposure (previous abortion) from cases and controls by means of personal interviews and by looking through national medical records. When interviewed, fewer controls admitted to having had an abortion than was evident in vital statistics. This discrepancy did not arise among cases. Differential recall between cases and controls led to a biased estimate of risk.

Bias from data gatherers presents further difficulties. If the individuals gathering information know the case or control status of the participants they can elicit information differently, again leading to potential information bias. A data gatherer might delve more deeply into a case's background than a control's to obtain a hypothesised exposure. When possible, data gatherers—eg, interviewers—should be unaware of the case or control status of the respondents. When blinding is not possible, investigators should keep the main hypothesis from the data gatherers. Furthermore, researchers should train data gatherers to elicit information similarly for cases and controls. Obtaining exposure information from records, as

Panel 2: Introduction of bias through poor choice of controls

Cases	Control selection	Non-representativeness	Selection bias
Colorectal cancer patients admitted to hospital	Patients admitted to hospital with arthritis	Controls probably have high degrees of exposure to NSAIDs	Would spuriously reduce the estimate of effect (odds ratio)
Colorectal cancer patients admitted to hospital	Patients admitted to hospital with peptic ulcers	Controls probably have low degrees of exposure to NSAIDs	Would spuriously increase the estimate of effect (odds ratio)

NSAIDs=non-steroidal anti-inflammatory drugs.

a solution to information bias, rarely suffices, since such information does not always exist, and, if it does, is usually insufficient to control adequately for confounding factors in the analysis.²²

Investigators who do case-control studies must be aware of the potential for information bias. They should address it in their study design and describe in their report approaches used to avoid such bias. Memory aids, such as photographs, diaries, and calendars can help participants remember exposures.²⁷ For example, in the case-control study of oral contraceptives,²⁷ the investigators used an album with colour photographs of every oral contraceptive marketed over the preceding decades and a blank calendar grid to help recall major life events and contraceptive use. Those colour photographs stimulated memories, both in cases and controls, to past exposure use, and thus reduced recall bias. Reports of case-control studies that do not detail use of memory aids, &c, should make readers sceptical.

Control for confounding

Case-control studies need to address confounding bias.^{21,22,29} This type of bias can be dealt with in the design phase by restriction or matching, but researchers generally prefer to handle it in the analysis phase with analytical techniques such as logistic regression or stratification with Mantel-Haenszel approaches.^{21,22,25} If this second approach is used, investigators should plan carefully in advance what potentially confounding variables to obtain data for; irrespective of the analytical approach used, researchers cannot control for a variable for which they have no data. Moreover, invalid measurement of potential confounding factors leads to residual confounding, even after adjustment.²²

Conclusion

Case-control studies that are well designed and carefully done can provide useful and reliable results. Investigators must, however, devote meticulous attention to the selection of control groups and to measurement of exposure information. Awareness of these key elements should help readers to identify the strengths and weaknesses of a properly reported study. Accurate and thorough description of methods by investigators will result in reader confidence in their results.

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of this material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

- Fuller Torrey E, Rawlings R, Yolken RH. The antecedents of psychoses: a case-control study of selected risk factors. *Schizophr Res* 2000; **46**: 17–23.
- Silverman DT. Risk factors for pancreatic cancer: a case-control study based on direct interviews. *Teratog Carcinog Mutagen* 2001; **21**: 7–25.
- Osaki Y, Minowa M. Factors associated with earthquake deaths in the great Hanshin-Awaji earthquake, 1995. *Am J Epidemiol* 2001; **153**: 153–56.
- Avidan B, Sonnenberg A, Schnell TG, Sontag SJ. Risk factors for erosive reflux esophagitis: a case-control study. *Am J Gastroenterol* 2001; **96**: 41–46.
- Freni-Titulaer LW, Kelley DB, Grow AG, McKinley TW, Arnett FC, Hochberg MC. Connective tissue disease in southeastern Georgia: a case-control study of etiologic factors. *Am J Epidemiol* 1989; **130**: 404–09.
- Strom BL, Reidenberg MM, West S, Snyder ES, Freundlich B, Stolley PD. Shingles, allergies, family medical history, oral contraceptives, and other potential risk factors for systemic lupus erythematosus. *Am J Epidemiol* 1994; **140**: 632–42.
- Amal NM, Lye MS, Ksiazek TG, et al. Risk factors for Nipah virus transmission, Port Dickson, Negeri Sembilan, Malaysia: results from a hospital-based case-control study. *Southeast Asian J Trop Med Public Health* 2000; **31**: 301–06.
- Traverso HP, Bennett JV, Kahn AJ, et al. Ghee applications to the umbilical cord: a risk factor for neonatal tetanus. *Lancet* 1989; **1**: 486–88.
- Cheng KK, Day NE, Duffy SW, Lam TH, Fok M, Wong J. Pickled vegetables in the aetiology of oesophageal cancer in Hong Kong Chinese. *Lancet* 1992; **339**: 1314–18.
- Hernandez J, Hawkins DL, Scollay MC. Race-start characteristics and risk of catastrophic musculoskeletal injury in thoroughbred racehorses. *J Am Vet Med Assoc* 2001; **218**: 83–86.
- Friedman GD, Hiatt RA, Quesenberry CP, Selby JV. Case-control study of screening for prostatic cancer by digital rectal examinations. *Lancet* 1991; **337**: 1526–29.
- Jick H, Zornberg GL, Jick SS, Seshadri S, Drachman DA. Statins and the risk of dementia. *Lancet* 2000; **356**: 1627–31.
- Cramer DW, Harlow BL, Titus-Ernstoff L, Bohlke K, Welch WR, Greenberg ER. Over-the-counter analgesics and risk of ovarian cancer. *Lancet* 1998; **351**: 104–07.
- Ingram D, Sanders K, Kolybaba M, Lopez D. Case-control study of phyto-oestrogens and breast cancer. *Lancet* 1997; **350**: 990–94.
- Collins JW, Smith GS, Baker SP, Landsittel DP, Warner M. A case-control study of forklift and other powered industrial vehicle incidents. *Am J Ind Med* 1999; **36**: 522–31.
- Wen LM, Estcourt CS, Simpson JM, Mindel A. Risk factors for the acquisition of genital warts: are condoms protective? *Sex Transm Infect* 1999; **75**: 312–16.
- Verloop J, Rookus MA, van der Kooy K, van Leeuwen FE. Physical activity and breast cancer risk in women aged 20–54 years. *J Natl Cancer Inst* 2000; **92**: 128–35.
- Slattery ML, Edwards SL, Ma KN, Friedman GD. Colon cancer screening, lifestyle, and risk of colon cancer. *Cancer Causes Control* 2000; **11**: 555–63.
- Abdel-Salam G, Czeizel AE. A case-control etiologic study of microcephaly. *Epidemiology* 2000; **11**: 571–75.
- Naghavi M, Barlas Z, Siadaty S, Naguib S, Madjid M, Casscells W. Association of influenza vaccination and reduced risk of recurrent myocardial infarction. *Circulation* 2000; **102**: 3039–45.
- Kelsey JL, Whittemore AS, Evans AS, Thompson WD. Methods in observational epidemiology. New York: Oxford University Press, 1996.
- Rothman KJ. Modern epidemiology. Boston: Little, Brown and Company, 1986.
- Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. *Lancet* 2002; **359**: 57–61.
- Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. *Lancet* 2002; **359**: 341–45.
- Schlesselman J. Case-control studies: design, conduct, analysis. New York: Oxford University Press, 1982.
- Moss AR, Osmond D, Bacchetti P, Chermann JC, Barre-Sinoussi F, Carlson J. Risk factors for AIDS and HIV seropositivity in homosexual men. *Am J Epidemiol* 1987; **125**: 1035–47.
- Stadel BV, Rubin GL, Webster LA, Schlesselman JJ, Wingo PA. Oral contraceptives and breast cancer in young women. *Lancet* 1985; **2**: 970–73.
- Lindfors-Harris BM, Eklund G, Adami HO, Meirik O. Response bias in a case-control study: analysis utilizing comparative data concerning legal abortions from two independent Swedish studies. *Am J Epidemiol* 1991; **134**: 1003–08.
- Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet* 2002; **359**: 248–52.

Generation of allocation sequences in randomised trials: chance, not choice

Kenneth F Schulz, David A Grimes

The randomised controlled trial sets the gold standard of clinical research. However, randomisation persists as perhaps the least-understood aspect of a trial. Moreover, anything short of proper randomisation courts selection and confounding biases. Researchers should spurn all systematic, non-random methods of allocation. Trial participants should be assigned to comparison groups based on a random process. Simple (unrestricted) randomisation, analogous to repeated fair coin-tossing, is the most basic of sequence generation approaches. Furthermore, no other approach, irrespective of its complexity and sophistication, surpasses simple randomisation for prevention of bias. Investigators should, therefore, use this method more often than they do, and readers should expect and accept disparities in group sizes. Several other complicated restricted randomisation procedures limit the likelihood of undesirable sample size imbalances in the intervention groups. The most frequently used restricted sequence generation procedure is blocked randomisation. If this method is used, investigators should randomly vary the block sizes and use larger block sizes, particularly in an unblinded trial. Other restricted procedures, such as urn randomisation, combine beneficial attributes of simple and restricted randomisation by preserving most of the unpredictability while achieving some balance. The effectiveness of stratified randomisation depends on use of a restricted randomisation approach to balance the allocation sequences for each stratum. Generation of a proper randomisation sequence takes little time and effort but affords big rewards in scientific accuracy and credibility. Investigators should devote appropriate resources to the generation of properly randomised trials and reporting their methods clearly.

“... having used a random allocation, the sternest critic is unable to say when we eventually dash into print that quite probably the groups were differentially biased through our predilections or through our stupidity.”¹

Until recently, investigators shunned formally controlled experimentation when designing trials (panel 1).²⁻⁵ Now, however, the randomised controlled trial sets the methodological standard of excellence in medical research (panel 2).^{3,6} The unique capability of randomised controlled trials to reduce bias depends on investigators being able to implement their principal bias-reducing technique—randomisation. Although random allocation of trial participants is the most fundamental aspect of a controlled trial,⁷ it unfortunately remains perhaps the least understood.^{8,9}

In this article, we describe the rationale behind random allocation and its related implementation procedures. Randomisation depends primarily on two interrelated but separate processes—ie, generation of an unpredictable randomised allocation sequence and concealment of that sequence until assignment occurs (allocation concealment). Here, we focus on how such a sequence can be generated. In a subsequent article, we will address allocation concealment.

What to look for with sequence generation

Non-random methods masquerading as random

Ironically, many researchers have decidedly non-random impressions of randomisation.⁸⁻¹⁰ They often mistake haphazard approaches and alternate assignment approaches as random.¹¹ Some medical researchers even

Lancet 2002; **359**: 515–19

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (K F Schulz PhD, D A Grimes MD)

Correspondence to: Dr Kenneth F Schulz (e-mail: KSchulz@fhi.org)

Panel 1: History of randomised controlled trials

The controlled trial gained increasing recognition during the 20th century as the best approach for assessment of health care and prevention alternatives. R A Fisher² developed randomisation as a basic principle of experimental design in the 1920s, and used the technique predominantly in agricultural research. The successful adaptation of randomised controlled trials to health care took place in the late 1940s, largely because of the advocacy and developmental work of Sir Austin Bradford Hill (figure) while at the London School of Hygiene and Tropical Medicine.³ His efforts culminated in the first experimental⁴ and published⁵ use of random numbers to allocate trial participants. Soon after, randomisation emerged as crucial in securing unbiased comparison groups.

Rights were not granted to include this image in electronic media. Please refer to the printed journal.

Austin Bradford Hill (1954)

The Wellcome Library, London

Panel 2: Benefits of randomisation

Proper implementation of a randomisation mechanism affords at least three major advantages:

It eliminates bias in treatment assignment

Comparisons of different forms of health interventions can be misleading unless investigators take precautions to ensure that their trial comprises unbiased comparison groups relative to prognosis. In controlled trials of prevention or treatment, randomisation produces unbiased comparison groups by avoiding selection and confounding biases. Consequently, comparison groups are not prejudiced by selection of particular patients, whether consciously or not, to receive a specific intervention. The notion of avoiding bias includes eliminating it from decisions on entry of participants to the trial, as well as eliminating bias from the assignment of participants to treatment, once entered. Investigators need to properly register each participant immediately on identification of eligibility for the trial, but without knowledge of the assignment. The reduction of selection and confounding biases underpins the most important strength of randomisation. Randomisation prevails as the best study design for study of small or moderate effects.⁶

It facilitates blinding (masking) of the identity of treatments from investigators, participants, and assessors, including the possible use of a placebo⁷

Such manoeuvres reduce bias after random assignment, and would be difficult, perhaps even impossible, to implement if investigators assigned treatments by a non-random scheme.

It permits the use of probability theory to express the likelihood that any difference in outcome between treatment groups merely indicates chance

view approaches antithetical to randomisation, such as assignment to intervention groups based on preintervention tests, as quasirandom.¹² Quasirandom, however, resembles quasipregnant, in that they both elude definition. Indeed, anything short of proper randomisation opens limitless contamination possibilities. Without properly done randomisation, selection and confounding biases seep into trials.^{7,13}

Researchers sometimes cloak, perhaps unintentionally, non-random methods in randomised clothing. They think that they have randomised by a method that, when described, is obviously not random. Methods such as assignment based on date of birth, case record number, date of presentation, or alternate assignment are not random, but rather systematic occurrences. Yet in a study that we did,¹⁰ in 5% (11 of 206) of reports investigators claimed that they had randomly assigned participants by such non-random methods. Furthermore, non-random methods are probably used much more frequently than suggested by our findings, since 63% (129 of 206) of the reports did not specify the method used to generate a random sequence.¹⁴

Systematic methods do not qualify as randomisation methods for theoretical and practical reasons. For example, in some populations, the day of the week on which a child is born is not entirely a matter of chance.¹⁵ Furthermore, systematic methods do not result in allocation concealment. By definition, systematic allocation usually precludes adequate concealment, since it results in previous knowledge of treatment assignment among those who recruit participants to the trial. If researchers report the use of systematic allocation, especially if masqueraded as randomised, readers should be wary of the results, since such a mistake implies

ignorance of the randomisation process. We place more credence in the findings of such a study if the authors accurately report it as non-randomised and explain how they controlled for confounding factors. In such instances, researchers should also discuss the degree of potential selection and information biases, allowing readers to properly judge the results in view of the non-random nature of the study and its biases.

Method of generation of an allocation sequence

To minimise bias, participants in a trial should be assigned to comparison groups based on some chance (random) process. Investigators use many different methods of randomisation,^{16–20} the most predominant of which are described.

Simple (unrestricted) randomisation

Elementary yet elegant describes simple randomisation (panel 3).²¹ Although the most basic of allocation approaches, analogous to repeated fair coin-tossing, this method preserves complete unpredictability of each intervention assignment. No other allocation generation approach, irrespective of its complexity and sophistication, surpasses the unpredictability and bias prevention of simple randomisation.²²

The unpredictability of simple randomisation, however, can also be a disadvantage.²³ With small sample sizes, simple randomisation (one-to-one allocation ratio) can yield highly disparate sample sizes in the groups by chance alone. For example, with a total sample size of 20, about 10% of the sequences generated with simple randomisation would yield a ratio imbalance of three to seven or worse.²⁴ This difficulty is diminished as the total sample size grows. Probability theory ensures that in the long term, the sizes of the treatment groups will not be greatly imbalanced. For a two-arm trial, the chance of pronounced imbalance becomes negligible with trial sizes greater than 200.²⁴ However, interim analyses with sample sizes of less than 200 might result in disparate group sizes.

Coin-tossing, dice-throwing, and dealing previously shuffled cards represent reasonable approaches for generation of simple complete randomisation sequences. All these manual methods of drawing lots theoretically lead to random allocation schemes, but frequently become non-random in practice. Distorted notions of randomisation

Panel 3: Simple randomisation

An almost infinite number of methods can be used to generate a simple randomisation sequence based on a random-number table.²¹ For example, for equal allocation to two groups, predetermine the direction to read the table: up, down, left, right, or diagonal. Then select an arbitrary starting point—ie, first line, 7th number:

```
56 99 20 20 52 49 05 78 58 50 62 86 52 11 88
31 60 26 13 69 74 80 71 48 73 72 18 60 58 20
55 59 06 67 02 ...
```

For equal allocation, an investigator could equate odd and even numbers to interventions A and B, respectively. Therefore, a series of random numbers 05, 78, 58, 50, 62, 86, 52, 11, 88, 31, &c, represent allocation to intervention A, B, B, B, B, B, A, A, B, A, &c. Alternatively, 00–49 could equate to A and 50–99 to B, or numbers 00–09 to A and 10–19 to B, ignoring all numbers greater than 19. Any of a myriad of options suffice, provided the assignment probabilities and the investigator adhere to the predetermined scheme.

sabotage the best of intentions. Fair coin-tossing, for example, allocates randomly with equal probability to two intervention groups, but can tempt investigators to alter the results of a toss or series of tosses—eg, when a series of heads and no tails are thrown. Many investigators do not really understand probability theory, and they perceive randomness as non-random. For example, the late Chicago baseball announcer Jack Brickhouse used to claim that when a 0.250 hitter (someone who would have a successful hit a quarter of the time) strolled to the plate for the fourth time, having failed the previous three times, that the batsman was “due”—ie, that the hitter would surely get a hit. However, Jack’s proclamation “he is due” portrayed a non-random interpretation of randomness. Similarly, a couple who have three boys and want a girl often think that their fourth child will certainly be a girl, yet the probability of them actually having a girl is still about 50%.

A colleague regularly demonstrated distorted views of randomisation with his graduate school class. He would have half his class develop allocation schemes with a proper randomisation method, and get the other half to develop randomisation schemes based on their personal views of randomisation. The students who used a truly random method would frequently have long consecutive runs of one treatment or the other. Conversely, students who used their own judgment would not. Class after class revealed their distorted impressions of randomisation.

Moreover, manual methods of drawing lots are more difficult to implement and cannot be checked. Because of threats to randomness, difficulties in implementation, and lack of an audit trail, we recommend that investigators avoid use of coin-tossing, dice-throwing, or card-shuffling, despite them being acceptable methods. Whatever method is used, however, should be clearly indicated in a researcher’s report. If no such description is made, readers’ should treat the study results with caution. Readers should have the most confidence in a sequence generation approach if the authors mention referral to either a table of random numbers or a computer random number generator, since these options represent unpredictable, reliable, easy, reproducible approaches that provide an audit trail.

Restricted randomisation

Restricted randomisation procedures control the probability of obtaining an allocation sequence with an undesirable sample size imbalance in the intervention groups.²⁰ In other words, if researchers want treatment groups of equal sizes, they should use restricted randomisation.

Blocking

Balanced (restricted) randomisation strives for unbiased comparison groups, but also strives for comparison groups of about the same size throughout the trial.²³ That attribute becomes helpful when investigators plan interim analyses. The use of simple randomisation might, upon occasion, produce quite disparate sample sizes at early interim analyses. Blocking obviates that problem.

The most frequently used method of achieving balanced randomisation is by random permuted blocks (blocking). For example, with a block size of six, of every six consecutively enrolled participants, three will normally be allocated to one treatment group and three to the other. However, the allocation ratio can be uneven. For example, a block size of six with a two-to-one ratio assigns four to one treatment group and two to the other in each block. This method can easily be extended to more than two treatments.

With blocking, the block size can remain fixed throughout the trial or be randomly varied. Indeed, if blocked randomisation is used in a trial that is not double-blinded, the block size should be randomly varied to reduce the chances of the assignment schedule being seen by those responsible for recruitment of participants.¹⁷ If the block size is fixed, especially if small (six participants or less), the block size could be deciphered in a not double-blinded trial. With treatment allocations becoming known after assignment, a sequence can be discerned from the pattern of past assignments. Some future assignments could then be accurately anticipated, and selection bias introduced, irrespective of the effectiveness of allocation concealment. Longer block sizes—eg, ten or 20—rather than smaller block sizes—four or six—and random variation of block sizes help preserve unpredictability.¹⁷

Investigators who do randomised controlled trials frequently use blocking. Those who report simply that they blocked, however, should make readers sceptical. Researchers should explicitly report having used blocking, the allocation ratio (usually one-to-one), the random method of selection (for example, random number table or computer random number generator), and the block size (or sizes if randomly varied).

Random allocation rule

The random allocation rule is the simplest form of restriction. For a particular total sample size, it ensures equal sizes only at the end of the trial. Usually, investigators identify a total sample size and then randomly choose a subset of that sample to assign to group A; the remainder are assigned to group B. For example, for a total study size of 200, placing 100 group A balls and 100 group B balls in a hat and drawing them randomly without replacement symbolises the random allocation rule. The sequence generation would randomly order 100 group A and 100 group B assignments. This method represents one large permuted-block for the entire study, which means that balance would usually only arise at the end of the trial and not throughout.

The random allocation rule maintains many of the positive attributes of simple complete randomisation, especially for statistical analysis, but is more likely to yield a chance covariate imbalance (chance confounding). It is noteworthy that this difference becomes trivial with larger sample sizes.²⁴ Moreover, unpredictability suffers compared with simple complete randomisation. Particularly in a non-double-blinded trial, scope exists for introduction of selection bias through guessing of assignments (especially toward the end of the trial), but obviously not at the level of permuted-block randomisation with small block sizes.^{24,25}

Investigators sometimes apply the random allocation rule by the restricted shuffled approach, which involves identifying the sample size, apportioning a number of specially prepared cards for each treatment according to the allocation ratio, inserting the cards into envelopes, and shuffling them to produce a form of random assignment without replacement.²³ Many investigators probably use this approach, but rarely call it restricted shuffled or the random allocation rule. Instead, they report use of envelopes or shuffling. Indeed, the restricted shuffled approach integrates, and conflates, allocation generation and concealment. Shuffling determines the allocation sequence, which is not optimum. Most importantly, the adequacy of the restricted shuffled approach pivots on proper allocation concealment with envelopes.^{7,8}

Biased coin and urn randomisation

Biased-coin designs achieve much the same objective as blocking but without forcing strict equality.^{16,26} They therefore preserve most of the unpredictability associated with simple randomisation. Biased-coin designs alter the allocation probability during the course of the trial to rectify imbalances that might be happening (panel 4). Adaptive bias-coin designs, with the urn design being the most widely studied, alter the probability of assignment based on the magnitude of the imbalance.

Biased-coin designs, including the urn design, appear infrequently in reports. They probably should, however, be used more often. Use of a computer is easier and more reliable than actually drawing balls from an urn, just as a computer is easier and more reliable than flipping a coin for simple randomisation. In unblinded trials, in which unpredictability becomes most important and the need for balance precludes simple randomisation, an urn design is especially useful. The unpredictability of urn designs surpasses permuted-block designs, irrespective of fixed or randomly varied block size approaches.²⁶ If readers encounter a biased-coin or urn design, they should consider it a proper sequence generation approach.

Replacement randomisation

Replacement randomisation repeats a simple randomisation allocation scheme until a desired balance is achieved. Trial investigators should establish objective criteria for replacement. For example, for a trial with 300 participants, investigators could specify that they would replace a simple randomisation scheme if the disparity between group sizes exceeds 20. If the first generated scheme's disparity exceeds 20, then they would generate an

Panel 4: Biased-coin and urn randomisation

Biased-coin approaches alter the allocation probability during the course of the trial to rectify imbalances in group numbers that might be happening. For example, investigators might use simple randomisation with equal probability of assignment—0.50/0.50 in a two-arm trial—as long as the disparity between the numbers assigned to the treatment groups remains below a prespecified limit. If the disparity reaches the limit, then investigators increase the probability of assignment to the group with the least participants—for example 0.60/0.40. Implemented properly, a biased-coin approach can achieve balance while preserving most of the unpredictability associated with simple randomisation.¹⁶

Adaptive bias-coin designs, with the urn design being the most widely studied, alter the probability of assignment based on the magnitude of the imbalance.²⁶ The urn design is designated as UD (α , β), with α being the number of blue and green balls initially and β representing the number of balls added to the urn of the opposite colour to the ball chosen (α and β being any reasonable non-negative numbers). For example in UD (2,1), an urn contains two blue balls and two green balls—0.50/0.50 probabilities to begin ($\alpha=2$). Balls are drawn at random and replaced for treatment assignments: blue for treatment A and green for treatment B. One additional ball ($\beta=1$) of the opposite colour to the ball chosen is added to the urn. If a blue ball was chosen first, then two blue balls and three green balls would be in the urn after the first assignment—0.40/0.60 for the next assignment. If another blue was chosen second, then two blue balls and four green balls would be in the urn after the second assignment—0.33/0.67 for the next assignment. That drawing procedure repeats with each assignment. The allocation probabilities fluctuate with the previous assignments.

entirely new simple randomisation scheme to replace the first attempt and check it against their objective criteria for disparity. They would iterate until they have a simple randomisation scheme that meets their criteria. Although replacement randomisation seems somewhat arbitrary, it is adequate as long as it is implemented before the trial begins. Moreover, it is easy to implement, ensures reasonable balance, and yields unpredictability. The main limitation of replacement randomisation is that it cannot ensure balance throughout the trial for interim analyses. Though rarely used, this approach emerged as the earliest form of restricted randomisation.^{20,27}

Stratified randomisation

Randomisation can create chance imbalances on baseline characteristics of treatment groups.²⁸ Investigators sometimes avert imbalances by use of prerandomisation stratification on important prognostic factors, such as age or disease severity. In such instances, researchers should specify the method of restriction (usually blocking). To reap the benefits of stratification, investigators must use a form of restricted randomisation to generate separate randomisation schedules for stratified subsets of participants defined by the potentially important prognostic factors. Stratification without restriction accomplishes nothing—ie, placebo stratification.

Stratification in trials is methodologically valid and useful, but theoretical and pragmatic issues limit its use to those planning new trials. The added complexity of stratification yields little additional gain in large trials, since randomisation creates balanced groups anyway. Moreover, if imbalance arises, then investigators can statistically adjust on those prognostic variables (preferably preplanned).^{28,29} Of greatest concern is that the added complexity of stratifying might discourage collaborators from participating in the trial or from entering participants during busy clinics, either of which affects enrolment. Thus, stratification in large trials offers negligible advantages coupled with important, pragmatic disadvantages. Note one important exception, however: stratification by centre in multicentre trials promises some benefit with no added complexity to the trial implementers within each centre. Also, another potential exception arises in large multicentre trials in which investigators use central randomisation for implementation of the sequence. Central randomisation limits the practical disadvantages of stratification and some gains might be realised in centres with smaller sample sizes.

Stratification might be useful in small trials in which it can avert severe imbalances on prognostic factors. It will confer adequate balance (on the stratified factors) and probably slightly more statistical power and precision.¹⁷ The gain from stratification becomes minimal, however, once the number of participants per group is more than 50.¹⁷ Moreover, stratification can indirectly cause negative effects if investigators seek exact balance within small strata. To achieve that exact balance, investigators often use small, fixed block sizes, which, in turn, hurts unpredictability.

Minimisation incorporates the general notions of stratification and restricted randomisation.¹⁶ It can be used to make small groups closely similar with respect to several characteristics. Minimisation in its strictest sense can be viewed as non-random,²¹ but, if used, we prefer a random component. Minimisation has supporters³⁰ and detractors.²² In any case, investigators who use minimisation should shield trial implementers from knowledge of upcoming assignments and other information that might facilitate guessing of upcoming assignments.¹⁶

Separation of generation and implementation

Investigators often neglect, usually unintentionally, one other important element of randomised controlled trial design and reporting. With all approaches, the people who generated the allocation scheme should not be involved in ascertaining eligibility, administering treatment, or assessing outcome. Such an individual would usually have access to the allocation schedule and thus the opportunity to introduce bias.⁸ Faults in this trial component might represent a crevice through which bias seeps into trials. Item ten (Implementation) in the CONSORT statement addresses this topic.^{28,31} Researchers should, therefore, state in reports who generated the allocation sequence, who enrolled participants, and who assigned participants. The person generating the allocation mechanism should be different from the person(s) enrolling and assigning. Nevertheless, under some circumstances, an investigator might have to generate the scheme and also enrol or assign. In such instances, the investigator should ensure the unpredictability of the assignment schedule and lock it away from everyone, particularly himself or herself.

Conclusion

Randomised controlled trials set the methodological standard of excellence in medical research. The key word is randomised, which must be done properly. Generation of a randomisation sequence takes little time and effort but affords big rewards in scientific accuracy and credibility. Investigators should devote appropriate resources to doing the generation properly and reporting their methods clearly.

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of the material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

- Hill AB. The clinical trial. *N Engl J Med* 1952; **247**: 133–19.
- Fisher RA. *The Design of Experiments*. London: Oliver and Boyo, 1935.
- Armitage P. The role of randomization in clinical trials. *Stat Med* 1982; **1**: 345–52.
- Doll R. Controlled trials: the 1948 watershed. *BMJ* 1998; **317**: 1217–20.
- Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *BMJ* 1948; **2**: 769–82.
- Peto R. Why do we need systematic overviews of randomized trials? *Stat Med* 1987; **6**: 233–44.
- Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273**: 408–12.
- Schulz KF. Subverting randomization in controlled trials. *JAMA* 1995; **274**: 1456–58.
- Schulz KF. Unbiased research and the human spirit: the challenges of randomized controlled trials. *CMAJ* 1995; **153**: 783–86.
- Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA* 1994; **272**: 125–28.
- Grimes DA. Randomized controlled trials: “it ain’t necessarily so”. *Obstet Gynecol* 1991; **78**: 703–04.
- Grimes D, Fraser E, Schulz K. Immunization as therapy for recurrent spontaneous abortion: a review and meta-analysis. *Obstet Gynecol* 1994; **83**: 637–38.
- Altman DG. Randomisation. *BMJ* 1991; **302**: 1481–82.
- Schulz KF, Chalmers I, Altman DG, Grimes DA, Dore CJ. The methodologic quality of randomization as assessed from reports of trials in specialist and general medical journals. *Online J Curr Clin Trials* 1995; Doc No 197: [81 paragraphs].
- MacFarlane A. Variations in number of births and perinatal mortality by day of week in England and Wales. *BMJ* 1978; **2**: 1670–73.
- Pocock SJ. *Clinical trials: a practical approach*. Chichester: Wiley, 1983.
- Meinert CL. *Clinical trials: design, conduct, and analysis*. New York: Oxford University Press, 1986.
- Friedman L, Furberg C, DeMets D. *Fundamentals of clinical trials*. St Louis: Mosby, 1996.
- Piantadosi S. *Clinical trials: a methodologic perspective*. New York: John Wiley and Sons, 1997.
- Lachin JM. Statistical properties of randomization in clinical trials. *Control Clin Trials* 1988; **9**: 289–311.
- Altman DG. *Practical statistics for medical research*. London: Chapman and Hall, 1991.
- Lachin JM, Matts JP, Wei LJ. Randomization in clinical trials: conclusions and recommendations. *Control Clin Trials* 1988; **9**: 365–74.
- Schulz KF. Randomized controlled trials. *Clin Obstet Gynecol* 1998; **41**: 245–56.
- Lachin JM. Properties of simple randomization in clinical trials. *Control Clin Trials* 1988; **9**: 312–26.
- Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. *Control Clin Trials* 1988; **9**: 327–44.
- Wei LJ, Lachin JM. Properties of the urn randomization in clinical trials. *Control Clin Trials* 1988; **9**: 345–64.
- Cox DR. *Planning of experiments*. New York: Wiley, 1958.
- Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; **134**: 663–94.
- Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I: introduction and design. *Br J Cancer* 1976; **34**: 585–612.
- Treasure T, MacRae KD. Minimisation: the platinum standard for trials? Randomisation doesn’t guarantee similarity of groups; minimisation does. *BMJ* 1998; **317**: 362–63.
- Moher D, Schulz KF, Altman DG, Lepage L. The CONSORT statement: revised recommendations for improving the quality of reports or parallel-group randomised trials. *Lancet* 2001; **357**: 1191–94.

Allocation concealment in randomised trials: defending against deciphering

Kenneth F Schulz, David A Grimes

Proper randomisation rests on adequate allocation concealment. An allocation concealment process keeps clinicians and participants unaware of upcoming assignments. Without it, even properly developed random allocation sequences can be subverted. Within this concealment process, the crucial unbiased nature of randomised controlled trials collides with their most vexing implementation problems. Proper allocation concealment frequently frustrates clinical inclinations, which annoys those who do the trials. Randomised controlled trials are anathema to clinicians. Many involved with trials will be tempted to decipher assignments, which subverts randomisation. For some implementing a trial, deciphering the allocation scheme might frequently become too great an intellectual challenge to resist. Whether their motives indicate innocent or pernicious intents, such tampering undermines the validity of a trial. Indeed, inadequate allocation concealment leads to exaggerated estimates of treatment effect, on average, but with scope for bias in either direction. Trial investigators will be crafty in any potential efforts to decipher the allocation sequence, so trial designers must be just as clever in their design efforts to prevent deciphering. Investigators must effectively immunise trials against selection and confounding biases with proper allocation concealment. Furthermore, investigators should report baseline comparisons on important prognostic variables. Hypothesis tests of baseline characteristics, however, are superfluous and could be harmful if they lead investigators to suppress reporting any baseline imbalances.

“The reason that the Medical Research Council’s controlled trial of streptomycin for pulmonary tuberculosis should be regarded as a landmark is thus not, as is often suggested, because random number tables were used to generate the allocation schedule . . . Rather it is because of the clearly described precautions that were taken to conceal the allocation schedule from those involved in entering patients.”¹

Generation of an unpredictable randomised allocation sequence represents the first crucial element of randomisation in a randomised controlled trial.² Implementation of the sequence, while concealing it at least until patients have been assigned to their groups (allocation concealment), is the important second element,^{3,4} without which, randomisation collapses in a trial.

As a direct consequence of randomisation, the first table in most reports of randomised controlled trials describes the baseline characteristics of the comparison groups.⁵ Researchers should describe their trial population and provide baseline comparisons of their groups so that readers can assess their comparability.⁵ In this article, we focus on proper approaches to allocation concealment and to reporting of baseline characteristics.

Allocation concealment

Researchers have many misconceptions with respect to allocation concealment. Proper allocation concealment secures strict implementation of a random allocation sequence without foreknowledge of treatment assignments. Allocation concealment refers to the technique used to implement the sequence,⁴ not to

generate it. Nevertheless, some people discuss allocation concealment with digressions into flipping coins or use of random number tables. Those digressions amount to methodological non-sequiturs; allocation concealment is distinct from sequence generation. Furthermore, some investigators confuse allocation concealment with blinding of treatments.^{3,4,6}

Without adequate allocation concealment, even random, unpredictable assignment sequences can be undermined.^{4,7,8} Knowledge of the next assignment could lead to the exclusion of certain patients based on their prognosis because they would have been allocated to the perceived inappropriate group. Moreover, knowledge of the next assignment could lead to direction of some participants to perceived proper groups, which can easily be accomplished by delaying a participant’s entry into the trial until the next appropriate allocation appears. Avoidance of such bias depends on the prevention of foreknowledge of treatment assignment. Allocation concealment shields those who admit participants to a trial from knowing the upcoming assignments. The decision to accept or reject a participant should be made, and informed consent should be obtained, in ignorance of the upcoming assignment.⁹

Importance of allocation concealment

Results of four empirical investigations^{4,10–12} have shown that trials that used inadequate or unclear allocation concealment, compared with those that used adequate concealment, yielded up to 40% larger estimates of effect. The badly done trials tended to exaggerate treatment effects. Moreover, the worst concealed trials yielded greater heterogeneity in results—ie, the results fluctuated extensively above and below the estimates from better studies.⁴ These findings provide empirical evidence that inadequate allocation concealment allows bias to seep into trials.

Indeed, having a randomised (unpredictable) sequence should make little difference without adequate allocation

Lancet 2002; **359**: 614–18

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (K F Schulz PhD, D A Grimes MD)

Correspondence to: Dr Kenneth F Schulz (e-mail: KSchulz@fhi.org)

concealment. Assume that investigators generate an adequate allocation sequence with a random number table. They then, however, post that sequence on a bulletin board, so that anyone involved in the trial could see the upcoming assignments. Similarly, the allocation sequence could be implemented through placing method indicator cards in translucent envelopes. This inadequate allocation concealment process could be deciphered by simply holding the envelopes to a bright light (figure). With both the bulletin board and the envelopes, those responsible for admitting participants could detect the upcoming treatment assignments and then channel individuals with a better prognosis to the experimental group and those with a poorer prognosis to the control group, or vice versa. Bias could easily be introduced, despite an adequate randomised sequence.⁷

Researchers should, therefore, ensure both adequate sequence generation and adequate allocation concealment in randomisation schemes.^{3,4,13} A mistake in either could compromise randomisation, resulting in incorrect results. For example, results of a trial could reveal a large treatment effect that only reflects a biased allocation procedure, or they could reveal no effect when in reality a harmful one prevails. Moreover, the results of such a trial can be more damaging than similar results from an explicitly observational research study.¹⁴ Biases are usually assumed and acknowledged in observational studies, and the statistical analysis and eventual interpretation attempt to take those biases into account. Conversely, studies labelled as randomised are frequently assumed to be free of bias, and commonly inadequate reporting masks the deficiencies they might have.^{3,13}

Consequently, the credibility of randomised controlled trials lends support to faster and greater changes in clinical or preventive management, which, if based on a compromised study, squanders scarce health resources, or even worse, harms peoples' health. Thus, the well-deserved credibility of randomised controlled trials produces an indirect liability. Wrong judgments emanate easily from improperly randomised trials.

Personal accounts of deciphering

Findings of empirical investigations^{4,10-12} suggest that investigators sometime undermine randomisation, though they rarely document such subversions. Nevertheless, when investigators responded anonymously to queries during epidemiological workshops, many did relate instances in which allocation schemes had been sabotaged.⁷

The individual accounts of such instances describe a range of simple to intricate operations.⁷ Most allocation concealment schemes were deciphered by investigators simply because the methods were inadequate. Investigators admitted, for instance, altering enrolment or allocations to particular study groups after decoding

future assignments, which were either posted on a bulletin board or visible through translucent envelopes held up to bright lights. Some also related opening unsealed assignment envelopes, sensing the differential weight of envelopes, or simply opening unnumbered envelopes until they found a desired treatment.

Investigators had a harder time deciphering the better allocation concealment schemes.⁷ Nevertheless, eventually someone described circumventing virtually every type of scheme. For example, some physicians took sequentially numbered, opaque, sealed envelopes to the hot light (an intense incandescent bulb) in the radiology department for deciphering of assignments. In studies using central randomisation, trial investigators related ringing the central number and asking for the next several assignments all at once; they received them in at least a couple of circumstances. In trials with sequentially numbered drug containers, someone described deciphering assignments based on the appearance of the container labels. Another had stopped trying to decipher a drug container scheme until she saw an attending physician, late at night, ransacking the office files of the principal investigator for the allocation list. Suggesting her methodological naïveté and innocence, she first thought of the attending physician's cleverness and not of the probability that such action would bias the trial.



Deciphering the allocation concealment scheme

Although investigators theoretically understand the need for unbiased research, they sometimes fail to maintain impartiality once they are involved in a trial. Researchers might want certain patients to benefit from one of the treatments, or the trial results to confirm their beliefs. Thus, certain trial procedures in properly done randomised controlled trials frustrate clinical inclinations, which annoys those doing the trial.^{7,15,16}

Some scientists aim to deliberately sabotage their results. However, many attempts at decoding the randomisation sequence simply indicate an absence of knowledge of the scientific ramifications of such actions. Furthermore, for some, the deciphering of the allocation scheme might frequently become too great an intellectual challenge to resist. As Oscar Wilde wrote, "The only way to get rid of temptation is to yield to it." Whether their motives are innocent or not, however, such tampering undermines the validity of a trial. Investigators must recognise the inquisitiveness of human nature and institute methodological safeguards. Proper allocation concealment will deter subversion, in effect, immunising trials against selection and confounding biases.^{7,15,16}

To develop a proper allocation scheme takes time, effort, and thought. Investigators cannot simply delegate this task without thoroughly examining the final product. Trial investigators will be crafty in any potential efforts to decipher the allocation sequence, so trial designers must be just as clever in their design efforts to prevent deciphering.

What to look for with allocation concealment

Researchers consider certain approaches to allocation concealment as adequate: sequentially numbered, opaque, sealed envelopes (SNOSE); pharmacy controlled; numbered or coded containers; central randomisation—eg, by telephone to a trials office—or other method whose description contained elements convincing of concealment—eg, a secure computer-assisted method.^{3,4,17} These criteria establish minimum methodological standards, yet they are met by only about a quarter of trials.^{3,17} Consequently, in assessment of allocation concealment from published reports, readers will be fortunate to find such standards reasonably met (panel 1).¹⁸⁻²³ Realistically, however, those minimum standards should be exceeded. If researchers provide descriptions that incorporate not only the minimum standards, but also elements of more rigorous standards, readers can have more confidence that selection and confounding biases have been averted (panel 2).

Methods that use envelopes are more susceptible to manipulation through human ingenuity than other approaches, and are therefore considered a less than ideal method of concealment.²⁴ If investigators use envelopes, they should diligently develop and monitor the allocation process to preserve concealment. In addition to use of sequentially numbered, opaque, sealed envelopes, they should ensure that the envelopes are numbered in advance, opened sequentially, and only after the participant's name and other details are written on the appropriate envelope.²⁵ We also recommend use of

pressure sensitive or carbon paper inside the envelope, which transfers such information to the assigned allocation and thus creates a valuable audit trail. Cardboard or aluminum foil placed inside the envelope further inhibits detection of assignments via hot lights.

Pharmacies can also engender both allocation-concealment and sequence-generation difficulties. Although reports in which the assignment was made by the pharmacy have generally been classified as having used an acceptable allocation concealment mechanism,^{3,4,17} compliance of pharmacists with proper randomisation methods in these trials is unknown. The precautions they took should have been reported. We are aware of instances in which pharmacists have violated assignment schedules.⁷ For instance, one large pharmacy charged a project US\$150 per participant for randomisation. During the course of the trial, over a weekend, the pharmacy ran out of one of the two drugs being compared, and therefore allocated the other drug to all newly enrolled participants to avoid slowing recruitment. We are aware of another pharmacy that randomised patients by alternate assignment. Investigators should not assume that pharmacists, and others involved in their trials, know about the methods of randomised controlled trials. Investigators must ensure that their research partners adhere to proper trial procedures. Beyond the minimum criteria, readers would gain additional confidence if investigators indicate that they instructed or checked the allocation mechanism of the pharmacy.

The use of sequentially numbered containers prevents foreknowledge of treatment assignment, but only if investigators take proper precautions. Beyond the minimum criteria, authors of trial reports should specify further details of the methods. Assurances that all of the

Panel 1: Descriptions of allocation concealment

“ . . . that combined coded numbers with drug allocation. Each block of ten numbers was transmitted from the central office to a person who acted as the randomisation authority in each centre. This individual (a pharmacist or a nurse not involved in care of the trial patients and independent of the site investigator) was responsible for allocation, preparation, and accounting of trial infusion. The trial infusion was prepared at a separate site, then taken to the bedside nurse every 24 h. The nurse infused it into the patient at the appropriate rate. The randomisation schedule was thus concealed from all care providers, ward physicians, and other research personnel.”¹⁸

“ . . . concealed in sequentially numbered, sealed, opaque envelopes, and kept by the hospital pharmacist of the two centres.”¹⁹

“Treatments were centrally assigned on telephone verification of the correctness of inclusion criteria . . .”²⁰

“Glenfield Hospital Pharmacy Department did the randomisation, distributed the study agents, and held the trial codes, which were disclosed after the study.”²¹

“The various placebo and treatment blocks were then issued with a medication number and assigned to consecutive patients in a sequential order. Two copies of the randomisation list were prepared: one was used by the packaging department, . . . supplied in blister packs containing 20 capsules for morning and evening administration over 10 days. These blister packs were supplied in labeled boxes—ie, one box for each patient and each dose.”²²

“Individuals were randomised by a computer-generated list, which was maintained centrally so no centre knew the treatment allocation of any patient. Marked capsule containers were designated for each patient, with additional containers being available should an increase to 15 mg or 20 mg sibutramine or placebo be prescribed by the centre's physician.”²³

Panel 2: Minimum and expanded criteria for adequate allocation concealment schemes

Minimum description of adequate allocation concealment scheme	Additional descriptive elements that provide greater assurance of allocation concealment
Sequentially numbered, opaque, sealed envelopes (SNOSE)	Envelopes are opened sequentially only after participant details are written on the envelope. Pressure-sensitive or carbon paper inside the envelope transfers that information to the assignment card (creates an audit trail). Cardboard or aluminum foil inside the envelope renders the envelope impermeable to intense light.
Sequentially numbered containers	All of the containers were tamper-proof, equal in weight, and similar in appearance.
Pharmacy controlled	Indications that the researchers developed, or at least validated, a proper randomisation scheme for the pharmacy. Indications that the researchers instructed the pharmacy in proper allocation concealment.
Central randomisation	The mechanism for contact—eg, telephone, fax, or e-mail—the stringent procedures to ensure enrolment before randomisation, and the thorough training for those individuals staffing the central randomisation office.

containers were tamper-proof, equal in weight, and similar in appearance, and that some audit trail had been established (such as writing the names of participants on the empty bottles or containers) would help readers to assess whether randomisation was likely to have been concealed successfully. Similarly, although central randomisation continues to be an excellent allocation concealment approach, effective trial procedures need to be established and followed. Researchers should at least specify the mechanism for contact—eg, telephone, fax, or e-mail—the stringent procedures to ensure enrolment prior to randomisation, and the thorough training of individuals at the central randomisation office. All these details should be addressed when doing a trial and when writing a trial report.^{7,13}

Other methods might suffice for adequate allocation concealment. Readers should look for descriptions that contain elements convincing of concealment. For example, a secure computer-assisted method might enable allocation concealment by preservation of assignments until enrolment is assured and confirmed. Indeed, automated assignment systems are likely to become more common.^{26,27} However, a simple computer system that merely stores assignments or naïvely shields assignments could turn out to be as transparent as tacking a randomisation list to a bulletin board. In describing an allocation concealment mechanism, investigators should display knowledge of the rationale behind allocation concealment and how their method met the standards.

Researchers frequently fail to report even the barest of descriptions of allocation concealment, preventing readers from assessing randomised controlled trials. The mechanism used to allocate interventions was omitted in reports of 93% of trials in dermatology,²⁸ 89% of trials in rheumatoid arthritis,²⁹ 48% of trials in obstetrics and gynecology journals,³ and 45% of trials in general medical journals.¹⁷ Fortunately, the situation is improving, since more medical journals are adopting reporting standards for randomised controlled trials.^{5,13,30} Moreover, with that reporting impetus, more investigators might design and do sound trials.

Baseline comparisons

Although randomisation eliminates systematic bias, it does not necessarily produce perfectly balanced groups with respect to prognostic factors. Differences due to chance remain in the intervention groups—ie, chance maldistribution. Statistical tests, however, account for these chance differences. The process of randomisation underlies significance testing and is independent of prognostic factors, known and unknown.³¹

Nevertheless, researchers should present distributions of baseline characteristics by treatment group in a table (table). Such information describes the hypothetical population from which their trial arose and allows readers to see the possibilities of generalisation to other populations.³² Furthermore, it allows physicians to infer the results to particular patients.⁵

	Antibiotic group (n=116)	Placebo group (n=129)
Characteristic		
Age (mean [SD]) (years)	30.2 (5.2)	31.1 (5.9)
Weight (median [25th, 75th centiles]) (kg)	141 (122, 181)	144 (123, 188)
Nulliparous (number, %)	62 (53%)	63 (49%)
Previous pelvic inflammatory disease (number, %)	24 (21%)	28 (22%)

Example of a reasonably reported table of baseline characteristics

A table of baseline characteristics also allows readers to compare the trial groups at baseline on important demographic and clinical characteristics. The common, inappropriate use of hypothesis tests—eg, p values in the tables—to compare characteristics concerns us, however.^{3,17,33,34} Such tests assess the probability that differences observed could have happened by chance. In properly randomised trials, however, any observed differences have, by definition, occurred by chance. “Such a procedure is clearly absurd,” as Altman states.³⁴

Hypothesis tests on baseline characteristics might not only be unnecessary but also harmful. Researchers who use hypothesis tests to compare baseline characteristics report fewer significant results than expected by chance.^{3,17} One plausible explanation for this discrepancy is that some investigators might have decided not to report significant differences, believing that by withholding that information they would increase the credibility of their reports. Not only are hypothesis tests superfluous, but they can be harmful if they indirectly lead investigators to suppress reporting baseline imbalances.

What to look for with baseline characteristics

Investigators should report baseline comparisons on important prognostic variables. Readers should look for comparisons based on consideration of the prognostic strength of the variables measured and the magnitude of any chance imbalances that have occurred, rather than statistical significance tests at baseline.³⁴ A table provides an efficient format of presenting baseline characteristics (table). Researchers should present continuous variables, such as age and weight, with an average and a measure of variability; usually a mean and standard deviation. If the data distribute asymmetrically, however, a median and a percentile range—ie, interquartile range—would provide better descriptions. Variability should not be described by standard errors and confidence intervals, since they are inferential rather than descriptive statistics.⁵ Numbers and proportions should be reported for categorical variables.⁵

In the analysis, the statistical tests on the outcomes account for any chance imbalances. Nevertheless, controlling for chance imbalances, if properly planned and done, might produce a more precise result.³⁵ Researchers should present any adjusted analyses and describe how and why they decided to adjust for certain covariates.

Conclusion

Proper randomisation remains the only way to avoid selection and confounding biases. The crucial unbiased nature of randomised controlled trials paradoxically coincides with their most vexing implementation problems. Randomised controlled trials antagonise human beings by frustrating their clinical inclinations. Thus, many involved with trials will be tempted to undermine randomisation, if afforded the opportunity to decipher assignments. To minimise the effect of this human tendency, trialists must devote meticulous attention to concealment of allocation schemes. Proper randomisation hinges on adequate allocation concealment.

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of this material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

- Chalmers I. Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments. *Int J Epidemiol* 2001; **30**: 1156–64.

- 2 Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. *Lancet* 2002; **359**: 515–19.
- 3 Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA* 1994; **272**: 125–28.
- 4 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273**: 408–12.
- 5 Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; **134**: 663–94.
- 6 Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomised trials. *Ann Intern Med* 2002; **136**: 254–59.
- 7 Schulz KF. Subverting randomization in controlled trials. *JAMA* 1995; **274**: 1456–58.
- 8 Pocock S. Statistical aspects of clinical trial design. *Statistician* 1982; **31**: 1–18.
- 9 Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline, I: control of bias and comparison with large co-operative trials. *Stat Med* 1987; **6**: 315–28.
- 10 Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998; **352**: 609–13.
- 11 Kjaergard L, Villumsen J, Gluud C. Quality of randomised clinical trials affects estimates of intervention. Abstracts for Workshops and Scientific Sessions, 7th International Cochrane Colloquium.
- 12 Jüni P, Altman D, Egger M. Assessing quality of controlled trials. *BMJ* 2001; **323**: 42–46.
- 13 Moher D, Schulz KF, Altman DG, Lepage L. The CONSORT statement: revised recommendations for improving the quality of reports or parallel-group randomised trials. *Lancet* 2001; **357**: 1191–94.
- 14 Torgerson DJ, Roberts C. Understanding controlled trials: randomisation methods—concealment. *BMJ* 1999; **319**: 375–76.
- 15 Schulz KF. Unbiased research and the human spirit: the challenges of randomized controlled trials. *CMAJ* 1995; **153**: 783–86.
- 16 Schulz KF. Randomised trials, human nature, and reporting guidelines. *Lancet* 1996; **348**: 596–98.
- 17 Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. *Lancet* 1990; **335**: 149–53.
- 18 Bellomo R, Chapman M, Finfer S, Hickling K, Myburgh J. Low-dose dopamine in patients with early renal dysfunction: a placebo-controlled randomised trial. Australian and New Zealand Intensive Care Society (ANZICS) Clinical Trials Group. *Lancet* 2000; **356**: 2139–43.
- 19 Smilde TJ, van Wissen S, Wollersheim H, Trip MD, Kastelein JJ, Stalenhoef AF. Effect of aggressive versus conventional lipid lowering on atherosclerosis progression in familial hypercholesterolaemia (ASAP): a prospective, randomised, double-blind trial. *Lancet* 2001; **357**: 577–81.
- 20 Anon. Low-dose aspirin and vitamin E in people at cardiovascular risk: a randomised trial in general practice. Collaborative Group of the Primary Prevention Project. *Lancet* 2001; **357**: 89–95.
- 21 Brightling CE, Monteiro W, Ward R, et al. Sputum eosinophilia and short-term response to prednisolone in chronic obstructive pulmonary disease: a randomised controlled trial. *Lancet* 2000; **356**: 1480–85.
- 22 McKeith I, Del Ser T, Spano P, et al. Efficacy of rivastigmine in dementia with Lewy bodies: a randomised, double-blind, placebo-controlled international study. *Lancet* 2000; **356**: 2031–36.
- 23 James WP, Astrup A, Finer N, et al. Effect of sibutramine on weight maintenance after weight loss: a randomised trial. STORM Study Group. Sibutramine Trial of Obesity Reduction and Maintenance. *Lancet* 2000; **356**: 2119–25.
- 24 Meinert CL. Clinical trials: design, conduct, and analysis. New York: Oxford University Press, 1986.
- 25 Bulpitt C. Randomised controlled clinical trials. Netherlands: Martinus Nijhoff, 1983.
- 26 Dorman K, Saade GR, Smith H, Moise KJ Jr. Use of the world wide web in research: randomization in a multicenter clinical trial of treatment for twin-twin transfusion syndrome. *Obstet Gynecol* 2000; **96**: 636–39.
- 27 Haag U. Technologies for automating randomized treatment assignment in clinical trials. *Drug Inform J* 1998; **118**: 7–11.
- 28 Adetugbo K, Williams H. How well are randomized controlled trials reported in the dermatology literature? *Arch Dermatol* 2000; **136**: 381–85.
- 29 Gotzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Control Clin Trials* 1989; **10**: 31–56.
- 30 Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* 2001; **285**: 1992–95.
- 31 Fisher RA. The design of experiments. Edinburgh: Oliver and Boyd, 1966.
- 32 Lachin JM. Statistical properties of randomization in clinical trials. *Control Clin Trials* 1988; **9**: 289–311.
- 33 Senn S. Statistical issues in drug development. Chichester: John Wiley and Sons, 1997.
- 34 Altman D. Comparability of randomised groups. *Statistician* 1985; **34**: 125–36.
- 35 Lavori PW, Louis TA, Bailar JC 3rd, Polansky M. Designs for experiments: parallel comparisons of treatment. *N Engl J Med* 1983; **309**: 1291–99.

Blinding in randomised trials: hiding who got what

Kenneth F Schulz, David A Grimes

Blinding embodies a rich history spanning over two centuries. Most researchers worldwide understand blinding terminology, but confusion lurks beyond a general comprehension. Terms such as single blind, double blind, and triple blind mean different things to different people. Moreover, many medical researchers confuse blinding with allocation concealment. Such confusion indicates misunderstandings of both. The term blinding refers to keeping trial participants, investigators (usually health-care providers), or assessors (those collecting outcome data) unaware of the assigned intervention, so that they will not be influenced by that knowledge. Blinding usually reduces differential assessment of outcomes (information bias), but can also improve compliance and retention of trial participants while reducing biased supplemental care or treatment (sometimes called co-intervention). Many investigators and readers naively consider a randomised trial as high quality simply because it is double blind, as if double-blinding is the sine qua non of a randomised controlled trial. Although double blinding (blinding investigators, participants, and outcome assessors) indicates a strong design, trials that are not double blinded should not automatically be deemed inferior. Rather than solely relying on terminology like double blinding, researchers should explicitly state who was blinded, and how. We recommend placing greater credence in results when investigators at least blind outcome assessments, except with objective outcomes, such as death, which leave little room for bias. If investigators properly report their blinding efforts, readers can judge them. Unfortunately, many articles do not contain proper reporting. If an article claims blinding without any accompanying clarification, readers should remain sceptical about its effect on bias reduction.

The rich history of blinding in clinical trials spans a couple of centuries.¹ Most researchers worldwide appreciate its meaning. Unfortunately, beyond that general appreciation lurks confusion. Terms such as single-blind, double-blind, and triple-blind mean different things to different people.² Moreover, many medical researchers confuse the term blinding with allocation concealment. The fact that such confusion arises suggests that both terms are misunderstood. Clear theoretical and practical differences separate the two. Blinding prevents ascertainment bias and protects the sequence after allocation.^{3,4} By contrast, researchers use methods of allocation concealment primarily to prevent selection bias and to protect an assignment sequence before and until allocation. Furthermore, in some trials, blinding cannot be successfully implemented, whereas allocation concealment can always be successfully implemented.^{4,5}

Blinding represents an important, distinct aspect of randomised controlled trials.³ The term blinding refers to keeping trial participants, investigators (usually health-care providers), or assessors (those collecting outcome data) unaware of an assigned intervention, so that they are not influenced by that knowledge. Blinding prevents bias at several stages of a trial, although its relevance varies according to circumstance. Although initial forays into blinding might have used a blindfold,¹ the processes have now become much more elaborate. In this article, we focus on the attributes and benefits of blinding.

Potential effects of blinding

If participants are not blinded, knowledge of group assignment can affect responses to the intervention

Lancet 2002; **359**: 696–700

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (K F Schulz PhD, D A Grimes MD)

Correspondence to: Dr Kenneth F Schulz (e-mail: KSchulz@fhi.org)

received.³ Participants who know that they have been assigned to a group who will receive a new treatment might harbour favourable expectations or increased apprehension. Those assigned to standard treatment, however, might feel deprived or relieved. Despite evidence to suggest that new treatments are as likely to be worse as they are to be better than standard treatments,⁶ participants probably assume that new treatments will be better than standard treatments—new means improved. In any case, knowledge of the intervention received, and perceptions of that treatment, can affect the psychological or physical responses of the participants. Knowledge of treatment allocation can also affect compliance and retention of trial participants (panel 1).

Blinding investigators—those who contribute to a broadly defined trial team including, but not limited to, trial designers, participant enrollers, randomisation implementors, health-care providers, intervention counsellors, and routine data collectors—is also important.³ Investigators especially pertinent to blinding include health-care providers (such as an attending physician or nurse) and intervention counsellors—eg, someone who delivers a behavioural prevention message—who might interact with the participants throughout the trial. If investigators are not blinded, their attitudes for or against an intervention can be directly transferred to participants.⁷ Their inclinations could also be manifested in differential use of ancillary interventions of supplemental care or treatment (co-interventions), differential decisions to withdraw participants from a trial, or differential adjustments to the medication dose (panel 1). Investigators might also encourage or discourage continuation in a trial on the basis of knowledge of the intervention group assignment.

Perhaps most importantly, blinding helps to reduce differential assessment of outcomes (often called information or ascertainment bias) (panel 1). For example, if outcome assessors who know of the treatment allocation believe a new intervention is better than an old

Panel 1: Potential benefits accruing dependent on those individuals successfully blinded

Individuals blinded	Potential benefits
Participants	<ul style="list-style-type: none"> Less likely to have biased psychological or physical responses to intervention More likely to comply with trial regimens Less likely to seek additional adjunct interventions Less likely to leave trial without providing outcome data, leading to lost to follow-up
Trial investigators	<ul style="list-style-type: none"> Less likely to transfer their inclinations or attitudes to participants Less likely to differentially administer co-interventions Less likely to differentially adjust dose Less likely to differentially withdraw participants Less likely to differentially encourage or discourage participants to continue trial
Assessors	<ul style="list-style-type: none"> Less likely to have biases affect their outcome assessments, especially with subjective outcomes of interest

one they could register more generous responses to that intervention. Indeed, in a placebo-controlled trial in patients with multiple sclerosis⁸ the unblinded, but not the blinded, neurologists' assessments showed an apparent benefit of the intervention.

Subjective outcomes—eg, pain scores—present great opportunities for bias.³ Furthermore, some outcomes judged objective can be fraught with subjectivity, for example, salpingitis. In general, though, blinding becomes less important to reduce observer bias as the outcomes become less subjective, since objective (hard) outcomes leave little opportunity for bias. Knowledge of the intervention would not greatly affect measurement of a hard outcome, such as death.

Lexicon of blinding

Non-blinded (open or open label) denotes trials in which everyone involved knows who has received which interventions throughout the trial. Blinding (masking) indicates that knowledge of the intervention assignments is hidden from participants, trial investigators, or assessors.

The terminology single blind usually means that one of the three categories of individuals (normally participant rather than investigator) remains unaware of intervention assignments throughout the trial.⁹ A single-blind trial might also, confusingly, mean that the participant and investigator both know the intervention, but that the assessor remains unaware of it.

In a double-blind trial, participants, investigators, and assessors usually all remain unaware of the intervention assignments throughout the trial.³ In view of the fact that three groups are kept ignorant, the terminology double blind is sometimes misleading. In medical research, however, an investigator frequently also assesses, so in this instance the terminology accurately refers to two categories.

Triple blind usually means a double-blind trial that also maintains a blind data analysis.¹⁰ Some investigators, however, denote trials as triple-blind if investigators and assessors are distinct people and both, as well as participants, remain unaware of assignments. Investigators rarely use quadruple blind, but those that do use the term to denote blinding of participants, investigators, assessors, and data analysts.¹¹ Thus, quintuple blind must mean that the allocation schedule has been lost and nobody knows anything. Contrary to Mae West's claim that "too much of a good thing can be wonderful", such is not always the case in blinding.

Confused terminology of single, double, and triple blinding permeates the literature,³ with physicians, textbooks, and journal articles all offering different interpretations and definitions.² Not only do investigators

not define double-blind trials consistently, in particular, but they make matters worse by frequently failing to report their definitions clearly in their articles. Building on the original blindfolding efforts,¹ and the once common double blindfold terminology,¹² we further obfuscate by offering additional definitions of single and double blinding (figure 1). More seriously, when we use double-blind or its derivatives in this article, we mean that steps have been taken to blind participants, investigators, and assessors to group assignments. In reporting randomised controlled trials, we urge researchers to explicitly state what steps they took to keep whom blinded.

Sparse reporting on blinding, however, is common. Many investigators neglect to report whether or not their trial was blinded. For example, reports of 51% of 506 trials in cystic fibrosis,¹³ 33% of 196 trials in rheumatoid arthritis,¹⁴ and 38% of 68 trials in dermatology¹⁵ did not state whether blinding was used. When researchers have reported their study as double-blind, they frequently have not provided much further clarification.^{14,16-18} For example, of 31 double-blind trials in obstetrics and gynecology, only 14 (45%) reports indicated the similarity of the treatment and control regimens (for example, appearance, taste, administration) and only 5 (16%) provided statements to indicate that blinding was successful.¹⁸

Masking or blinding

Some people prefer the term masking to blinding to describe the same procedure. Masking might be more appropriate in trials that involve participants who have impaired vision, and could be less confusing in trials in which blindness is an outcome.³ Blinding, however, conveys a strong bias prevention message. Apparently, blinding terminology emerged when Benjamin Franklin and colleagues¹⁹ actually blindfolded participants to shield them from knowledge in their assessments of the



Figure 1: The authors: double blinded versus single blinded



Figure 2: **The authors blinded and masked**

therapeutic claims made for Mesmerism. The imagery of blindfolding, a total covering of the eyes, conveys stronger bias prevention than masking, where eye holes could permit viewing (figure 2). Blinding also suggests a more secure procedure to some. The International Conference on Harmonization (ICH) guidance,²⁰ for example, primarily uses blinding terminology. (The ICH is an intensive tripartite collaboration between regulatory authorities in Europe, Japan, and the USA to develop common guidelines for the design, implementation, and reporting of clinical trials). We prefer blinding because it has a long history, maintains worldwide recognition, creates strong imagery, and permeates the ICH guidelines.³

Placebos and blinding

Interventions (treatments) sometimes have no effect on the outcomes being studied.³ When an ineffective intervention is administered to participants in the context of a well-designed randomised controlled trial, however, beneficial effects on participants' attitudes sometimes occur, which in turn affect outcomes.¹⁰ Researchers refer to this phenomena as the placebo effect.

A placebo refers to a pharmacologically inactive agent that investigators administer to participants in the control group of a trial.³ The use of a placebo control group balances the placebo effect in the treatment group, allowing for independent assessment of the treatment effect. Although placebos can have a psychological effect, they are administered to participants in a trial because they are otherwise inactive. An active placebo is a placebo with properties that mimic the symptoms or side-effects—eg, dry mouth, sweating—that might otherwise reveal the identity of the (pharmacologically) active test treatment. Most researchers agree that placebos should be administered, whenever possible, to controls when assessing the effects of a proposed new treatment for a condition for which no effective treatment already exists.^{9,10} Indeed, blinding frequently necessitates the use of placebos.

However, a proven effective standard treatment, if such exists, is usually given to the control group for comparison against a new treatment.³ Thus, investigators might compare two active treatment groups without a placebo group. Even then, however, investigators frequently attempt to achieve blinding by use of the double-dummy method, in essence two placebos.^{11,21} For example, for comparison of two agents, one in a blue capsule and the other in a red capsule, the investigators would prepare blue placebo capsules and red placebo capsules. Then both treatment groups would receive a blue and a red capsule, one active and one inactive.

Does blinding prevent bias?

Some investigators, readers, and editors overstate the importance of blinding in prevention of bias. Indeed, some consider a randomised trial as high quality if it is double blind—ie, as if double blinding is the sine qua non of a randomised controlled trial.³ Unfortunately, scientific life is not that simple. A randomised trial can be methodologically sound and not be double blind or, conversely, double blind and not methodologically sound. Lasagna¹² captured that notion long ago: “Let us examine the placebo somewhat more critically, however, since it and ‘double blind’ have reached the status of fetishes in our thinking and literature. The Automatic Aura of Respectability, Infallibility, and Scientific Savoir-faire which they possess for many can be easily shown to be undeserved in certain circumstances.”¹² Although double blinding suggests a strong design, it is not the primary indicator of overall trial quality. Moreover, many trials cannot be double blinded. Such trials must, therefore, be judged on overall merit rather than an inapplicable standard based on double blinding.

We do not, however, suggest that blinding is unimportant.³ Intuitively, blinding should reduce bias, and available evidence supports that impression. Methodological investigations tend to show that double blinding prevents bias but is less important, on average, in prevention of bias than is adequate allocation concealment.^{4,22,23}

What to look for in descriptions of blinding

In general, if researchers describe a trial as double-blind, readers can assume that they have avoided bias. Empirical evidence lends support to this recommendation. As suggested in the CONSORT guidelines,^{24,25} however, investigators should not use only the single-blind, double-blind, or triple-blind terminology, but should also explicitly state who was blinded, and how. Moreover, if the researchers contend that the trial investigators, participants, and assessors were blinded—ie, double blind—then they should provide information about the mechanism (capsules, tablets, film, &c), similarity of treatment characteristics (appearance, taste, administration), and allocation schedule control—eg, location of the schedule during the trial, when the code was broken for the analysis, and circumstances under which the code could be broken for individual instances. Such additional information can lend support to or undermine claims of double-blinding (panel 2).^{26–29}

If researchers properly report their blinding efforts, readers can judge those efforts. Unfortunately, many articles will not contain proper reporting. If a researcher claims to have done a blinded study, but does not provide accompanying clarification, readers should remain sceptical about its effect on bias reduction. For example, one trial³⁰ of prophylactic antibiotics claimed to be blinded, but the methods section of the report revealed that little or no blinding occurred.

Ideally, researchers should also relate if blinding was successful. Investigators can theoretically assess the success of blinding by directly asking participants, health-care providers, or outcome assessors which intervention they think was administered (panel 3). In principle, if blinding was successful, these individuals should not be able to do better than chance when guessing the intervention, for example. In practice, however, blinding might be totally successful, but participants, health-care providers, and outcome assessors might nevertheless guess the intervention because of ancillary information. Disproportionate levels of adverse side-effects might

Panel 2: Descriptions of blinding

"No patient, research nurse, investigator, or any other medical or nursing staff in the ICU was aware of the treatment assignments for the duration of the study. All statistical analysis was also done with masking maintained.

Randomisation authorities were instructed to report any suspected breach of the masking procedures. No report was filed . . . The drug or placebo (vehicle without active drug) was prepared for syringe pump infusion or for volumetric pump infusion in indistinguishable syringes or bags."²⁶

". . . in a double-blind, placebo-controlled manner . . . Neither the patients nor doctors could distinguish the placebo from sibutramine capsules. The taste of the capsules was identical provided they were swallowed whole as instructed. . . Results of biochemical analyses were completed before the randomisation code was broken at the end of the completed trial."²⁷

"The study was double-blinded—that is, neither the women nor the study staff, including the biostatisticians at Family Health International, knew which group was using the nonoxynol 9 film. The nonoxynol 9 film contained . . . The placebo film contained . . . The films were identical in appearance, packaging, and labeling."²⁸

"The doxycycline and placebo were in capsule form and identical in appearance . . . The randomization code was kept in the USA." (Note: the trial was conducted in Kenya) "Thus, all administration and assessments were done blinded to treatment assignment, and the investigators and patients were also blinded to the ongoing results of the study. The code was broken only after data collection had been completed."²⁹

ICU=intensive care unit.

provide strong hints as to the intervention. Irrespective of painstaking efforts to do double-blinded trials, some interventions have side-effects that are so recognisable that their occurrence will unavoidably reveal the intervention received to both the participants and the health-care providers.^{11,24} Even more fundamental than hints from adverse effects are the hints from clinical outcomes. Researchers usually welcome large clinical effects (except perhaps in equivalence trials). If they arise, health-care providers and participants would likely deduce—not always accurately of course—that a participant with a positive outcome received the active (new) intervention rather than control (standard). If indeed the active (new) intervention materialises as helpful (highly desirable) then their deductions would be correct more often than chance guesses.^{24,31} Irrespective of their suspicions, end-of-trial tests of blindness might actually be tests of hunches for adverse effects or efficacy.^{32,33}

Furthermore, individuals might be reluctant to expose any unblinding efforts by providing accurate responses to

Panel 3: Assessment of the success of blinding

"We asked 126 staff members their opinions of which film was the placebo. Eighteen percent thought film A (the placebo) was the placebo, 13 percent thought film B (nonoxynol 9) was the placebo, and 69 percent had no opinion about which film was the placebo. Of the 68 peer educators (the staff members most likely to reflect the opinion of the participants), 16 percent thought film A was the placebo, 13 percent thought film B was the placebo, and 71 percent had no opinion."²⁸

the queries—in other words, if they have deciphered group assignments, they might provide responses contrary to their deciphering findings to disguise their actions. That difficulty, along with interpretation difficulties stemming from adverse side-effects and successful clinical outcomes, leads us to question the usefulness of tests of blinding in some circumstances. Investigators should carefully consider the usefulness of assessing the success of their blinding efforts, but if they proceed, should provide the results of any assessments. At the very least, they should report any failure of the blinding procedure, such as non-identical placebo or active preparations. Published reports rarely contain assessments of blinding, but, if provided, readers should sceptically assess the information presented.

Double blinding proves difficult or impossible in many trials. For instance, in general, surgical trials cannot be double blinded. Specifically, a trial that compares degrees of pain associated with sampling blood from the ear or thumb cannot be double-blinded.³⁴ If researchers do not describe their trial as double-blind or the equivalent, it could still be scientifically strong. Apart from assessment of the other methodological aspects of the trial, readers would have to assess how much bias might have ensued due to absence of blinding. Readers should identify if anybody was blinded in the trial and what benefits might have accrued (panel 1). Indeed, blinding of outcome assessors is often possible and advisable, even in open trials.¹¹ For example, lesions can be photographed before and after treatment and assessed by someone not involved in the study.¹¹ We recommend placing greater credence in results when someone unaware of treatment assignments judges outcome measures.

Even that recommendation, however, is not absolute. As noted earlier, some hard outcomes, such as death, leave little room for ascertainment bias. In other words, blinding the assessor to hard outcomes might have little effect.

Conclusion

Blinding embodies a rich history spanning over two centuries. Most researchers worldwide understand blinding terminology, but confusion lurks beyond a general comprehension. Investigators should clearly explicate those blinded and not blinded in their trial, rather than only labeling their trial as single-blind, double-blind, or triple-blind. Readers should expect such clarity when reading and judging a trial report.

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of this material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

- 1 Kaptchuk TJ. Intentional ignorance: a history of blind assessment and placebo controls in medicine. *Bull Hist Med* 1998; **72**: 389–433.
- 2 Devereaux PJ, Manns BJ, Ghali WA, et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. *JAMA* 2001; **285**: 2000–03.
- 3 Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. *Ann Intern Med* 2002; **136**: 254–59.
- 4 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273**: 408–12.
- 5 Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA* 1994; **272**: 125–28.
- 6 Chalmers I. What is the prior probability of a proposed new treatment being superior to established treatments? *BMJ* 1997; **314**: 74–75.

- 7 Wolf S. Effects of suggestion and conditioning on action of chemical agents in human subjects: pharmacology of placebos. *J Clin Invest* 1950; **29**: 100–09.
- 8 Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology* 1994; **44**: 16–20.
- 9 Meinert CL. Clinical trials: design, conduct, and analysis. New York: Oxford University Press, 1986.
- 10 Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley, 1983.
- 11 Day SJ, Altman DG. Statistics notes: blinding in clinical trials and other studies. *BMJ* 2000; **321**: 504.
- 12 Lasagna L. The controlled trial: theory and practice. *J Chronic Dis* 1955; **1**: 353–67.
- 13 Cheng K, Smyth RL, Motley J, O’Hea U, Ashby D. Randomized controlled trials in cystic fibrosis (1966–1997) categorized by time, design, and intervention. *Pediatr Pulmonol* 2000; **29**: 1–7.
- 14 Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Control Clin Trials* 1989; **10**: 31–56.
- 15 Adetugbo K, Williams H. How well are randomized controlled trials reported in the dermatology literature? *Arch Dermatol* 2000; **136**: 381–85.
- 16 Mosteller F, Gilbert JP, McPeck B. Reporting standards and research strategies for controlled trials: agenda for the editor. *Controlled Clin Trials* 1980; **1**: 37–58.
- 17 DerSimonian R, Charette LJ, McPeck B, Mosteller F. Reporting on methods in clinical trials. *N Engl J Med* 1982; **306**: 1332–37.
- 18 Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *BMJ* 1996; **312**: 742–44.
- 19 Franklin B, Bailly JS, Lavoisier A. Rapport des commissaires chargés par le roi, de l’examen du magnétisme animal. Nice: Chez Gabriel Floteron, 1785.
- 20 Department of Health and Human Services, FDA. International conference on harmonisation: guidance on statistical principles for clinical trials. *Fed Regist* 1998; **63**: 49583–98.
- 21 Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991.
- 22 Khan KS, Daya S, Collins JA, Walter SD. Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertil Steril* 1996; **65**: 939–45.
- 23 Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998; **352**: 609–13.
- 24 Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; **134**: 663–94.
- 25 Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports or parallel-group trials. *Lancet* 2001; **357**: 1191–94.
- 26 Bellomo R, Chapman M, Finfer S, Hickling K, Myburgh J. Low-dose dopamine in patients with early renal dysfunction: a placebo-controlled randomised trial. Australian and New Zealand Intensive Care Society (ANZICS) Clinical Trials Group. *Lancet* 2000; **356**: 2139–43.
- 27 James WP, Astrup A, Finer N, et al. Effect of sibutramine on weight maintenance after weight loss: a randomised trial. STORM Study Group. *Lancet* 2000; **356**: 2119–25.
- 28 Roddy RE, Zekeng L, Ryan KA, Tamoufe U, Weir SS, Wong EL. A controlled trial of nonoxynol 9 film to reduce male-to-female transmission of sexually transmitted diseases. *N Engl J Med* 1998; **339**: 504–10.
- 29 Sinei SK, Schulz KF, Lamptey PR, et al. Preventing IUCD-related pelvic infection: the efficacy of prophylactic doxycycline at insertion. *Br J Obstet Gynaecol* 1990; **97**: 412–19.
- 30 Baker KR, Drutz HP, Barnes MD. Effectiveness of antibiotic prophylaxis in preventing bacteriuria after multichannel urodynamic investigations: a blind, randomized study in 124 female patients. *Am J Obstet Gynecol* 1991; **165**: 679–81.
- 31 Quitkin FM, Rabkin JG, Gerald J, Davis JM, Klein DF. Validity of clinical trials of antidepressants. *Am J Psychiatry* 2000; **157**: 327–37.
- 32 Sackett DL, Gent M, Taylor DW. Tests for the blindness of randomized trials may not. *Clin Res* 1986; **34**: 711A.
- 33 The Canadian Cooperative Study Group. A randomized trial of aspirin and sulfinpyrazone in threatened stroke. *N Engl J Med* 1978; **299**: 53–59.
- 34 Carley SD, Libetta C, Flavin B, Butler J, Tong N, Sammy I. An open prospective randomised trial to reduce the pain of blood glucose testing: ear versus thumb. *BMJ* 2000; **321**: 20.

Sample size slippages in randomised trials: exclusions and the lost and wayward

Kenneth F Schulz, David A Grimes

Proper randomisation means little if investigators cannot include all randomised participants in the primary analysis. Participants might ignore follow-up, leave town, or take aspartame when instructed to take aspirin. Exclusions before randomisation do not bias the treatment comparison, but they can hurt generalisability. Eligibility criteria for a trial should be clear, specific, and applied before randomisation. Readers should assess whether any of the criteria make the trial sample atypical or unrepresentative of the people in which they are interested. In principle, assessment of exclusions after randomisation is simple: none are allowed. For the primary analysis, all participants enrolled should be included and analysed as part of the original group assigned (an intent-to-treat analysis). In reality, however, losses frequently occur. Investigators should, therefore, commit adequate resources to develop and implement procedures to maximise retention of participants. Moreover, researchers should provide clear, explicit information on the progress of all randomised participants through the trial by use of, for instance, a trial profile. Investigators can also do secondary analyses on, for instance, per-protocol or as-treated participants. Such analyses should be described as secondary and non-randomised comparisons. Mishandling of exclusions causes serious methodological difficulties. Unfortunately, some explanations for mishandling exclusions intuitively appeal to readers, disguising the seriousness of the issues. Creative mismanagement of exclusions can undermine trial validity.

Proper randomisation^{1,2} means little if investigators cannot include all randomly assigned participants in their primary analysis. Hence, a crucial aspect of assessing a randomised controlled trial pertains to exclusions, withdrawals, losses, and protocol deviations. How should investigators handle participants who refuse entry, ignore follow-up, leave town, or take aspartame when they were instructed to take aspirin? Unfortunately, many inappropriate approaches to dealing with these types of problem actually seem logical and falsely appealing. Therein lies their insidious nature, because such inappropriate approaches can result in serious biases. Here, we address the effect of exclusions made before and after randomisation.

Exclusions before randomisation

Investigators can exclude participants before randomisation. The eventual randomised treatment comparison will remain unbiased (good internal validity), irrespective of whether researchers have well-founded or whimsical reasons for exclusion of particular individuals. However, exclusions at this stage can hurt extrapolation, the generalisability, of the results (external validity). For most investigations, we therefore recommend that eligibility criteria be kept to a minimum, in the spirit of the large and simple trial.^{3,4} However, some valid reasons exist for exclusion of certain participants. Individuals could, for example, have a condition for which an intervention is contraindicated, or they could be judged likely to be lost to follow-up. The trial question should guide the approach.⁵ Sometimes, however, investigators impose so many eligibility criteria that their trial infers to a population of little apparent interest to anyone, and, in addition, recruitment becomes difficult. If investigators exclude too many participants, or the wrong participants,

their results might not represent the people of interest, even though the randomised controlled trial might have been meticulously done—ie, the results could be true but potentially irrelevant.

What to look for in exclusions before randomisation

The eligibility criteria should indicate the population to which the investigators wish to infer. When judging the results of a trial, readers should make sure that the eligibility criteria are clear and specific. Most importantly, the criteria should have been applied before randomisation. Readers should also assess whether any of the criteria make the study sample atypical, unrepresentative, or irrelevant to the people of interest. In practice, however, results from a trial will infrequently be totally irrelevant: “most differences between our patients and those in trials tend to be quantitative (they have different ages or social classes or different degrees of risk of the outcome event or of responsiveness to therapy) rather than qualitative (total absence of responsiveness or no risk of the event).”⁶ Such qualitative differences in response are rare; thus, trials tend to have rather robust external validity.⁶

Exclusions after randomisation

Exclusions made after randomisation threaten to bias treatment comparisons. Randomisation itself configures unbiased comparison groups at baseline. Any erosion, however, over the course of the trial from those initially unbiased groups produces bias, unless, of course, that erosion is random, which is unlikely. Consequently, for the primary analysis, methodologists suggest that results for all patients who are randomly assigned should be analysed, and, furthermore, should be analysed as part of the group to which they were initially assigned.^{3,7} Trialists refer to such an approach as an intent-to-treat analysis. Simply put: once randomised, always analysed as assigned.

Intent-to-treat principles underlie the primary analysis in a randomised controlled trial to avoid biases associated with non-random loss of participants.^{8–10} Investigators can

Lancet 2002; **359**: 781–85

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (K F Schulz PhD, D A Grimes MD)

Correspondence to: Dr Kenneth F Schulz (e-mail: KSchulz@fhi.org)

also do secondary analyses, preferably preplanned, based on only those participants, for example, who fully comply with the trial protocol (per protocol) or who receive the treatment irrespective of randomised assignment (on-treatment or as-treated). Secondary analyses are acceptable as long as researchers label them as secondary and non-randomised comparisons. Trouble brews, however, when investigators exclude participants and, in effect, present a secondary, non-randomised comparison as the primary randomised comparison from a trial. In reality, this analysis represents a cohort study masquerading as a randomised controlled trial. Exclusion of participants from an analysis can lead to misleading conclusions (panel 1).^{11–14}

Researchers often do not provide adequate information on excluded participants.^{7,15,16} Furthermore, in a review of 249 randomised controlled trials published in major general medical journals in 1997, only 2% (five of 249) of reports explicitly stated that all randomly assigned participants were analysed according to the randomised group assignment.¹⁷ About half of the reports (119 of 249) noted an intent-to-treat analysis, but many provided no details to support this claim.

Additionally, researchers frequently do not report anything with respect to exclusions.⁷ Left in this information void, many readers deduce that certain trials used intent-to-treat principles and had no exclusions. We call this scenario no apparent exclusions. Readers commonly view trials with no apparent exclusions as less biased, when in fact unreported exclusions probably occurred in many of them. Indeed, trials with no apparent exclusions were methodologically weaker than those reporting at least some exclusions.⁷ In other words, some of the more biased trials might be mistakenly interpreted as unbiased, and many of the less biased trials as biased; we call this inconsistency the exclusion paradox. Until researchers comprehensively report exclusions after randomisation, readers should be aware of this unsettling irony.

What to look for in exclusions after randomisation

Before we launch into attributes of proper handling of exclusions after randomisation, we should acknowledge the tenuous ground on which any such discussion rests. Reporting on exclusions is poor, with the exclusion

paradox misleading readers. Investigators should provide clear, explicit information on the progress through the trial of all randomised participants, and when such information is absent, readers should be sceptical. The flow diagrams specified in the CONSORT statement provide appropriate guidelines.^{18,19}

Optimally, of course, investigators would have no exclusions after randomisation and use an intent-to-treat analysis. Assessment of exclusions after randomisation is simple: none are allowed. All participants enrolled should be analysed as part of the original group assigned. Clinical research is not normally that simple, but the principle holds. One pragmatic hint for minimising exclusions after randomisation involves randomly assigning individuals at the last possible moment. If randomisation takes place when the participant is first identified, but before treatment is initiated, then any exclusions arising before treatment still become exclusions after randomisation. Investigators can address this potential difficulty by delaying randomisation until immediately before treatment begins.²⁰

If investigators report exclusions after randomisation, those exclusions should be carefully scrutinised because they could bias comparisons. Exclusions arise after randomisation for several reasons, including discovery of patient ineligibility, postrandomisation-pretreatment outcome, deviation from protocol, and losses to follow-up.

Discovery of participant ineligibility

In some trials, participants are enrolled and later discovered not to have met the eligibility criteria. Exclusions at this point could seriously bias the results, since discovery is probably not random. For example, participants least responsive to treatment or who have side-effects might draw more attention and, therefore, might be more likely to be judged ineligible than other study participants. Alternatively, a physician who had treatment preferences for certain participants might withdraw individuals from the trial if they were randomly assigned to what he believes to be the wrong group.

Participants discovered to be ineligible should remain in the trial. An exception could be made if establishment of eligibility criteria is difficult. In such instances, investigators could obtain the same information from each patient at time of randomisation and have it centrally

Panel 1: A randomised controlled trial of sulfinpyrazone versus placebo for prevention of repeat myocardial infarction

For this trial, the researchers reported a primary analysis that compared rates of death from cardiac causes rather than from all cardiac deaths.^{11,12} In their analysis, inappropriate exclusions due to eventual discovery of patient ineligibility caused a problem:¹³ the investigators withdrew as ineligible seven patients who had received treatment—six in the treatment group and one in the placebo group—resulting in more patients who died being withdrawn from the treatment group than from the placebo group.

Moreover, results of a detailed audit of this trial by the US Food and Drugs Administration (FDA) indicate that additional patients from the placebo group could have been declared ineligible on the basis of similar criteria, but were not.¹³ Furthermore, the trial protocol did not mention exclusion of ineligible patients after entry, particularly patients who had died. The researchers also excluded two deaths in the sulfinpyrazone group and one death in the placebo group as non-analysable because of poor compliance. However, the trial protocol did not include plans to exclude patients because of poor compliance.

Additionally, the investigators used a 7-day rule. They declared as non-analysable any death of a patient who had not received treatment for at least 7 days or who died more than 7 days after termination of treatment. The FDA review committee did not criticise this practice strongly, principally because the protocol described the 7-day rule, and also because the rule had little overall effect on the results.

Overall, these inappropriate exclusions did, however, affect the results of the study.¹³ Although the researchers initially reported a 32% reduction ($p=0.058$) in rates of death from cardiac causes for participants who took the drug, a reanalysis showed a weaker result. When individuals judged ineligible or non-analysable were included in the originally assigned groups, the reduction was only 21% ($p=0.16$). It is noteworthy that only p values were provided. We urge the use of confidence intervals in reporting results.¹⁴ Moreover, the fallout from inappropriate exclusions, as ascertained by the FDA, cast doubt over the trial. The FDA advisory committee announced that sulfinpyrazone could not be labelled and advertised as a drug to prevent death in the critical months after a heart attack because, on close examination, the data were not as convincing as they seemed at first glance.

reviewed by an outside source, blinded to the assigned treatment. That source, whether a person or group, could then withdraw patients who did not satisfy the eligibility criteria, presumably in an unbiased way.

Postrandomisation, pretreatment outcome

Researchers sometimes report exclusion of participants on the basis of outcomes that happen before treatment has begun or before the treatment could have had an effect. For example, in a clinical trial of a specific drug's effect on death rates, investigators withdrew as non-analysable data on all patients who died after randomisation but before treatment began or before they had received at least 7 days of treatment. This winnowing seems intuitively attractive, because none of the deaths can then be attributable to treatment. But the same argument could be made for excluding data on all patients in a placebo group who died during the entire study interval, because, theoretically, none of these deaths could have been related to treatment. This example illustrates the potential for capriciousness in addressing postrandomisation, pretreatment outcomes.

Randomisation tends to balance the non-attributable deaths in the long run. Any tinkering after randomisation, even if done in the most scientific and impartial manner, cannot improve upon that attribute, but can hurt it. More importantly, this meddling sometimes serves as a post hoc rationale for inappropriate exclusions.

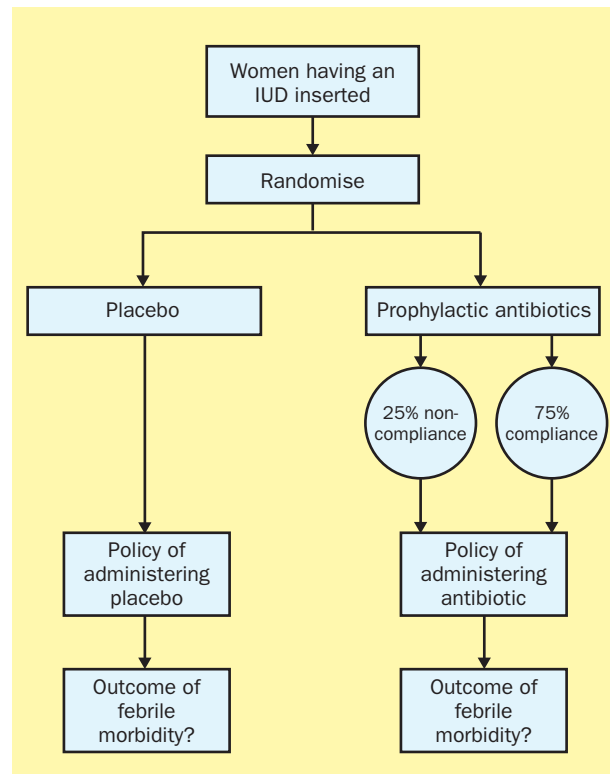
Post hoc rationalisation arises when investigators observe the results and then frame rules that favour their hypotheses. Assume that an investigator postulates that a drug used for treatment reduced the death rate associated with a particular condition. After analysis of data, however, the investigator notes that 14 deaths in the treatment group and two deaths in the placebo group arose before treatment had begun or before the drug had been taken for at least 7 days. She then rationalises the deaths as unrelated to treatment, and withdraws them from analysis. Such a response would seriously bias her results, even though her reasoning in the report would likely seem logical.

Imposed a priori, such rules only complicate trial implementation; imposed a posteriori, they lead to biased and invalid results. In assessment of randomised controlled trials, identification of when researchers stipulated rules usually proves impossible. We prefer to find, in reports of randomised controlled trials, that investigators did not allow any withdrawal of participants after randomisation. The data of all randomised patients should be analysed. Planned or unplanned, the exclusion of non-analysable outcomes on grounds of efficiency is not a generally accepted practice in the analysis of a randomised clinical trial.²¹

Protocol deviations

Deviations from assigned treatment happen in many trials. Some investigators suggest that participants who deviate substantially from the allotted treatment should be excluded in the final analysis, or should be included only up to the point of deviation. Although this approach seems attractive, it has a serious flaw: "the group which deviates from one protocol and the group which deviates from the other protocol may be so different [. . .] that the treatment comparison in the remaining patients will be severely biased."³

For example, suppose investigators want to know if prophylactic antibiotics reduce febrile morbidity associated with insertion of an intrauterine device (IUD). Investigators randomly allocate participants to receive



Schematic of randomised IUD patients, accounting for their compliance with treatment during the trial

IUD=intrauterine device.

antibiotics or placebo (figure). Unfortunately, 25% of the patients in the antibiotic group deviate from the protocol and do not take their antibiotics. In effect, these deviates receive the same treatment—that is, nothing—as the placebo group. Should the investigators exclude them from analysis? Alternatively, should investigators merge them with the placebo group and compare them with the compliant patients in the antibiotic group who adhered to the protocol? Some investigators opt for one of these speciously attractive options.

For the primary analysis, however, neither option proves acceptable. The two treatment groups would no longer be comparable. The participants who did not take antibiotics might have been in better health or might have better tolerated insertion of the IUD. In either instance, they were probably less susceptible to febrile morbidity. If investigators exclude the deviates, the antibiotic group will contain only the more susceptible: the treatment comparison would be even more biased. If investigators include the deviates in the placebo group, then not only will those left in the antibiotic group be more susceptible to febrile morbidity, but the placebo group will have been infiltrated with less susceptible patients: the treatment comparison would be even more biased. Those who deviated could be sicker rather than healthier—it does not matter. The point remains that the treatment comparison would be systematically biased.

All protocol deviations should be followed up, and their data should be analysed with the group to which they were originally assigned. In our example, the deviates from the antibiotic group should remain with the antibiotic group. Similarly, any deviates in the placebo group should remain in that group. Despite what happened during the course of the trial, investigators should compare the group randomly allocated to antibiotics with the group allocated

to placebo. This approach, in addition to being unbiased, will provide a pragmatic answer to the question of primary clinical interest—eg, does the policy of giving prophylactic antibiotics for IUD insertion prevent febrile morbidity? Thus, if researchers report excluding protocol deviates, or if they report moving protocol deviates from one group to another group, the resultant treatment comparison should be considered biased, analogous to an observational study.

Loss to follow-up

Losses to follow-up are perhaps the most vexing of the proffered reasons for exclusions after randomisation. Participants might move or might refuse to continue participating in the trial. Participants lost to follow-up could still be included in the analysis if outcome information could be obtained from another source, such as gathering data from a national death registry. Such opportunities, however, rarely arise. Without outcomes from those lost to follow-up, investigators have little choice but to exclude them from the analysis. Any losses damage internal validity, but differential rates of loss among comparison groups cause major damage. Hence, investigators must minimise their losses to follow-up.

Minimisation of loss in some trials exudes difficulties. Investigators should commit adequate attention and resources to develop and implement procedures to minimise losses.¹⁰ For example, investigators might exclude patients before randomisation if deemed likely to be lost to follow-up. Alternatively, they could obtain contact information to locate lost participants or hire special follow-up personnel who visit unresponsive participants, or both.

Some investigators add innovative twists that cultivate high follow-up rates. One approach uses a large number of conveniently placed follow-up clinics. Too often investigators expect participants to visit a single, inconvenient location. Shortening the data collection instrument to a manageable size caters to the participants' wishes and needs. Investigators foster follow-up by not overburdening participants. Such instruments might not only promote higher follow-up rates, but might also engender higher quality data on the main items of interest. Elimination of loss completely could be impossible, but investigators too frequently profess insurmountable difficulties. Many investigators could work harder than they do to obtain higher follow-up rates (panel 2).

What is an acceptable rate of loss to follow-up? Only one answer, 0%, ensures the benefits of randomisation. Obviously, this is unrealistic at times. Some researchers suggest a simple five-and-20 rule of thumb, with fewer than 5% loss probably leading to little bias, greater than 20% loss potentially posing serious threats to validity, and in-between levels leading to intermediate levels of problems.²² Indeed, in their experience with sensitivity analyses, use of the worst case scenario, they opine, and we agree, that a trial would be unlikely to successfully withstand challenges to its validity with losses of more than 20%.⁶ Indeed, some journals refuse to publish trials with losses greater than 20%.⁶

Although the five-and-20 rule is useful, it can oversimplify the problem in situations with infrequent outcomes.²² Expectations for loss to follow-up depend on various factors, such as the topic examined, the outcome event rate, and the length of follow-up. For example, if researchers examined outcomes during the first day after birth to women delivering in hospitals, we would expect no losses. If the researchers examined use of microbicides by women in Africa (who usually have no phones and

Panel 2: Approaches to maximisation of participant follow-up

Hire a person to manage and encourage follow-up

Hire personnel to call participants or to visit participants at their homes or place of work, if participants are not returning for follow-up

Exclude before randomisation those likely to be unwilling to return

Exclude before randomisation those likely to move

Obtain contact information to prompt participants to return for follow-up and to facilitate location of participants if they do not return—eg, mail, telephone, and e-mail for enrolled participants, for close friends or relatives who do not live with the participant, and for the participant's family doctor

Obtain an identification number, such as a national health-care number

Establish follow-up venues suited to participants rather than to investigators and trial implementers—eg, more locations than just the central clinic or hospital, close to where participants live, convenient to access, and sensitive to waiting time

Streamline trial procedures to move participants quickly through a follow-up visit

Keep data collection instrument short so as to not overburden the participant

Provide excellent and free medical care

Provide monetary subsidies, primarily for time and travel costs incurred by participants

sometimes lack street addresses) to prevent HIV-1 transmission over a 1-year follow-up period, however, we would expect perhaps 5–15% loss to follow-up, although hoping for lower. Actually, most investigators have done much worse under such circumstances, but recent exhaustive efforts have yielded loss to follow-up rates of about 1.5%.²³ Another useful general rule of thumb suggests not allowing the loss to follow-up rate to exceed the outcome event rate.

Perhaps more important than the absolute overall loss to follow-up rate is the comparative loss rates in the groups. Researchers should analyse the data for differential rates of loss in the groups. Bias could arise when losses are related to differences in unpleasantness, toxicity, or efficacy of the treatments. In any case, investigators should have recorded and analysed the outcomes from those participants lost, at least up to the point of loss.

Conclusion

Trialists should endeavour to minimise exclusions after randomisation and to do intent-to-treat analyses. They should also follow the CONSORT statement for reporting.^{18,19} The flow diagram (trial profile) helps particularly to track the progress of participants through a trial.

For readers, non-reporting of exclusions results in interpretation difficulties, such as the exclusion paradox, which misleads readers about trial quality. Moreover, mishandling of exclusions causes serious methodological difficulties. Unfortunately, some explanations provided in reports for such difficulties intuitively appeal to readers, which disguises the seriousness of the issues. Readers must battle both inadequate reporting and their intuition to discover potential threats to validity.

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of the material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

- Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. *Lancet* 2002; **359**: 515–19.
- Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. *Lancet* 2002; **359**: 614–18.
- Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient, I: introduction and design. *Br J Cancer* 1976; **34**: 585–612.
- Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient, II: analysis and examples. *Br J Cancer* 1977; **35**: 1–39.
- Sackett DL. On some prerequisites for a successful clinical trial. In: Shapiro SH, Louis TA, eds. *Clinical trials: issues and approaches*. New York: Marcel Dekker, 1983: 65–79.
- Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine: how to practice and teach EBM*. Edinburgh: Churchill Livingstone, 2000.
- Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *BMJ* 1996; **312**: 742–44.
- Lee YJ, Ellenberg JH, Hirtz DG, Nelson KB. Analysis of clinical trials by treatment actually received: is it really an option? *Stat Med* 1991; **10**: 1595–605.
- Lewis JA, Machin D. Intention to treat: who should use ITT? *Br J Cancer* 1993; **68**: 647–50.
- Lachin JM. Statistical considerations in the intent-to-treat principle. *Control Clin Trials* 2000; **21**: 167–89.
- The Anturane Reinfarction Trial Research Group. Sulfinpyrazone in the prevention of cardiac death after myocardial infarction. *N Engl J Med* 1978; **298**: 289–95.
- The Anturane Reinfarction Trial Research Group. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. *N Engl J Med* 1980; **302**: 250–56.
- Temple R, Pledger GW. The FDA's critique of the anturane reinfarction trial. *N Engl J Med* 1980; **303**: 1488–92.
- Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. *Lancet* 2002; **359**: 57–61.
- DerSimonian R, Charette LJ, McPeck B, Mosteller F. Reporting on methods in clinical trials. *N Engl J Med* 1982; **306**: 1332–37.
- Meinert CL, Tonascia S, Higgins K. Content of reports on clinical trials: a critical review. *Control Clin Trials* 1984; **5**: 328–47.
- Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 1999; **319**: 670–74.
- Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports or parallel-group trials. *Lancet* 2001; **357**: 1191–94.
- Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; **134**: 663–94.
- Friedman L, Furberg C, DeMets D. *Fundamentals of clinical trials*. St Louis: Mosby, 1996.
- Meier P. Anturane reinfarction trial. *N Engl J Med* 1981; **304**: 730.
- Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine: how to practice and teach EBM*. New York: Churchill Livingstone, 1997.
- Roddy RE, Zekeng L, Ryan KA, Tamoufé U, Tweedy KG. The effect of Nonoxynol-9 gel on urogenital gonorrhoea and chlamydial infection: randomized controlled trial. *JAMA* (in press).

Uses of error

No problem

Paolo Gallinaro

25 years ago, I started another orthopaedic night shift on duty. My colleague, finishing his, told me that a young man with a pelvic fracture had been admitted an hour or so earlier. “There is no problem with him”, he said, “his blood pressure is 120/80 mm Hg, and he has an intravenous line with saline running in.” Radiographs showed an open book pelvis: a fractured anterior ring and slight disruption of the sacroiliac joints. Immediate action did not seem necessary, so I took care of other patients with minor injuries. Only later did I have a look at the “no problem” patient. Although his blood

pressure had not changed, he was pale and his extremities were cold. Clearly, he had haemorrhagic shock. I called a general surgeon. He suspected an arterial retroperitoneal bleed, and decided to operate. Opening the large retroperitoneal haematoma just made the bleeding worse—and he found no major source. The patient died. Patients with similar injuries still die as a result of similar mistakes, but we have since changed the management, and prognosis, of closed pelvic fractures by using aggressive fluid resuscitation, stabilisation, and intensive care.

Dipartimento di Traumatologia, Ortopedia e Medicina del Lavoro, Università di Torino, 10126 Torino, Italy (Prof P Gallinaro MD)

Unequal group sizes in randomised trials: guarding against guessing

Kenneth F Schulz, David A Grimes

We cringe at the pervasive notion that a randomised trial needs to yield equal sample sizes in the comparison groups. Unfortunately, that conceptual misunderstanding can lead to bias by investigators who force equality, especially if by non-scientific means. In simple, unrestricted, randomised trials (analogous to repeated coin-tossing), the sizes of groups should indicate random variation. In other words, some discrepancy between the numbers in the comparison groups would be expected. The appeal of equal group sizes in a simple randomised controlled trial is cosmetic, not scientific. Moreover, other randomisation schemes, termed restricted randomisation, force equality by departing from simple randomisation. Forcing equal group sizes, however, potentially harms the unpredictability of treatment assignments, especially when using permuted-block randomisation in non-double-blinded trials. Diminished unpredictability can allow bias to creep into a trial. Overall, investigators underuse simple randomisation and overuse fixed-block randomisation. For non-double-blinded trials larger than 200 participants, investigators should use simple randomisation more often and accept moderate disparities in group sizes. Such unpredictability reflects the essence of randomness. We endorse the generation of mildly unequal group sizes and encourage an appreciation of such inequalities. For non-double-blinded randomised controlled trials with a sample size of less than 200 overall or within any principal stratum or subgroup, urn randomisation enhances unpredictability compared with blocking. A simpler alternative, our mixed randomisation approach, attains unpredictability within the context of the currently understood simple randomisation and permuted-block methods. Simple randomisation contributes the unpredictability whereas permuted-block randomisation contributes the balance, but avoids the perfect balance that can result in selection bias.

A tantalising phone call begins, “I have just read a report of a randomised trial, and I found problems!” All too often, however, the discussion proceeds with: “Look at that difference in sample sizes in the groups—they are not equal. I am suspicious of this trial.” Or in planning a trial, “What can we do to end up with equal sample sizes?” Indeed, large disparities in sample sizes not explained by chance should cause concern,^{1,2} but many researchers look askance at a trial with any disparity. We cringe at this seemingly ubiquitous notion that a randomised trial needs to yield equal sample sizes. Somehow such a notion seems imbedded in many a medical researcher’s psyche.

Such conceptual misunderstanding deters prevention of bias in trials. Exactly equal sample sizes in a randomised controlled trial contribute little to statistical power and potentially harm unpredictability, especially in non-double-blinded trials that use permuted-block randomisation. Unpredictability reflects the essence of randomisation because those involved cannot predict the next treatment assignment. With predictability comes bias.

Greater predictability emanates from randomisation schemes that depart from simple, unrestricted randomisation. Such departures are termed restricted randomisation schemes.^{3,4} They constrain treatment assignment schedules to yield similar or, most frequently, equal group sizes throughout the trial, assuming the most common desired allocation ratio of one-to-one. The restricted randomisation schemes all sacrifice unpredictability, but that increased predictability primarily surfaces in non-double-blinded trials that use permuted-blocks (panel 1).⁵⁻⁷

Trialists rely on the security of unpredictability. In the

past, we suggested cultivation of a tolerance for groups of unequal sample sizes in simple randomised trials.^{8,9} We now suggest cultivation of a tolerance for groups of unequal sizes in restricted randomisation trials as well.

Forcing cosmetic credibility

Studies reported as randomised yield equal sample sizes in the comparison groups more frequently than expected.⁸⁻¹⁰ In simple, unrestricted randomised controlled trials (analogous to repeated coin-tossing), the relative sizes of comparison groups should indicate random variation. In other words, some discrepancy between the numbers in the comparison groups would be expected. However, analyses of reports of trials in general and specialist medical journals showed that researchers too frequently reported equal sample sizes of the comparison groups (defined as exactly equal or as equal as possible in view of an odd number total sample size).^{8,9} In the specialist journals, the disparity of sample sizes in the comparison groups deviated from expected ($p < 0.001$) and produced equal group sizes in 54% of the simple randomised (unrestricted) trials.⁸ This result was higher than that in blocked trials (36%), and blocked trials aspire for equality. Moreover, results of a similar analysis of the dermatology literature showed that an even higher 71% of simple randomised trials reported essentially equal group sizes.¹⁰

Why would investigators seek equal or similar sample sizes in comparison groups? We feel many investigators strive for equal sample sizes as an end in itself. The lure of the so-called cosmetic credibility of equal sizes seems apparent. Sadly, that cosmetic credibility also appeals to readers. Striving for equal sample sizes with simple randomisation, however, reflects a methodological non sequitur.

The high proportion of equal group sizes noted above represent pronounced aberrations from chance occurrences and suggest non-random manipulations of assignments to force equality. Other logical explanations seem plausible, but probably do not account for the degree of aberration

Lancet 2002; **359**: 966–70

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (K F Schulz PhD, D A Grimes MD)

Correspondence to: Dr Kenneth F Schulz (e-mail: KSchulz@fhi.org)

Panel 1: Unpredictability in allocation sequences

Predictability in clinical trials breeds bias. If trial investigators identify or predict upcoming allocation assignments, they can instill selection bias. In assessment of eligibility, they could exclude a participant destined for, in their opinion, the wrong group. Moreover, various manoeuvres allow them to channel participants with a better prognosis to the experimental group and those with a poorer prognosis to the control group, or vice versa.^{5,6} Irrespective of the reasons for doing so, experimenters bias the comparison. Clinicians might revere predictability in caring for patients, but they must understand that predictability spawns bias in clinical trials.

Trial investigators can guess the next assignments by subverting the allocation concealment mechanism—eg, by holding translucent envelopes to a light bulb.^{5,7} However, proper allocation concealment usually prevents this subversion. Alternatively, with permuted-block randomisation, trial investigators can sometimes predict the next assignments by noting a pattern of past assignments.^{4,5} For example, in a non-double-blinded trial with a block size of four, if a trial investigator notes that the sample size in the two groups equilibrates after every four participants, then many future assignments can be predicted. For example, if the sequence ABA materialises in a block of four, B would necessarily be the next assignment, or if the sequence BB materialises, AA would be the next two assignments.

In non-double-blinded trials, all intervention allocations become known after assignment, even with proper allocation concealment. Thus, if a pattern to the allocation sequence exists, the trial investigator can discern it and predict some future assignments. However, if no pattern exists, or if the pattern is indiscernible, the allocation sequence is unpredictable. Therefore, knowledge of past assignments would not help in prediction of future assignments. Unpredictability is essential in non-double-blinded randomised trials.

Proper allocation concealment before assignment and proper blinding of all involved in the trial after assignment shields knowledge of past assignments and thereby prevents prediction of future assignments. Proper blinding diminishes the need for unpredictability. Even in supposedly blinded trials, however, blinding after assignment is not always successful. If trial investigators perceive quickly developing, clinically obvious side-effects that reveal the intervention assigned, for instance, blinding might not prevent predictions.

witnessed.^{8,10} Such tinkering with assignments creates difficulties by directly instilling selection bias into trials. We hope to remove some of those difficulties by dispelling the mythology behind the drive for exactly equal sizes.

Beyond the issue of non-random manipulations of assignments, however, we will concentrate on the potential bias introduced by balancing group sizes with valid restricted randomisation methods, primarily permuted-block randomisation, that produce equal group sizes throughout the trial. Unfortunately, methods used to ensure equal sample sizes can facilitate correct future predictions of treatment assignments, allowing bias to infiltrate.

Unequal group sizes in restricted trials

The method of restricted randomisation is used to balance sample sizes. That balance usually enhances statistical power and addresses any time trends that might exist in treatment efficacy and outcome measurement during the course of a trial.^{11,12} Moreover, restricted randomisation within strata becomes essential for investigators to attain the benefits of stratification.¹³ Thus, reasonable scientific justification lends support to restriction.

For restriction to be effective, however, it need not yield exactly equal sample sizes. The power of a trial is not sensitive to slight deviations from equality of the sample sizes.³ Thus, restricted approaches that produce similar sizes would yield power, time trend, and stratification benefits much the same as those restricted randomisation approaches that produce equal sizes.

Equal sample sizes, however, can have negative consequences. The predominant restricted randomisation method is random permuted-blocks (blocking). Such an approach effectively attains the goals of equal sample sizes in the comparison groups overall (and, if stratified, within strata). Moreover, the method generates equal sample sizes after every block. With that attribute, however, comes the disadvantage of predictability.^{5,11}

Predictability, particularly, becomes a major weakness in a non-double-blinded trial. We define a double-blinded trial as one in which the treatment is hidden from participants, investigators, and outcome assessors. In virtually all non-double-blinded trials, some investigators become aware of the treatment. Thus, even with adequate allocation concealment, treatment assignments become known after assignment. With that information, trial investigators can unravel the fixed block size (presumably the organisers

initially shielded all block size information from them) and then anticipate when equality of the sample sizes will arise (panel 1). A sequence can be discerned from the pattern of past assignments and then some future assignments could be accurately anticipated. Hence selection bias could seep in, irrespective of the effectiveness of allocation concealment.^{4,5} The same difficulty to a lesser degree might be true in a double-blinded trial in which obvious, perceptible side-effects materialise quickly.

Randomised controlled trials become prone to unravelling of block sizes when the block size remains fixed throughout the trial, especially if the block size is small—eg, six or fewer participants. Hence, if investigators use blocked randomisation, they should randomly vary the block size to lower the chances of an assignment schedule being inferred by those responsible for recruitment and assignment.^{4,5}

Random block sizes, however, are no panacea. Even with random variation of block sizes, blocking still generates equal sample sizes many times throughout a trial. Indeed, based on a modification of a model that measures inherent predictability of intervention assignments with certainty, random block sizes, at best, decrease but do not eliminate the potential for selection bias.¹¹ In other words, random block sizes help to reduce, but in some instances might not eliminate, selection bias. Permuted-block randomisation, even with random block sizes, presents trial recruiters with opportunities to anticipate some assignments.

Alternatives in non-double-blinded trials

For non-double-blinded randomised controlled trials with an overall sample size of more than 200 (an average sample size of 100 in two groups) and within each planned subgroup or stratum, we recommend simple randomisation.¹⁴ It provides perfect unpredictability thereby eliminating that aspect of selection bias due to the generation of the allocation sequence. Moreover, simple randomisation also provides the least probability for chance bias of all the generation procedures,³ and it enables valid use of virtually all standard statistical software. With sample sizes greater than 200, simple randomisation normally yields only mild disparities in sample sizes between groups. The cut-off of 200, however, is merely an overall guideline. Individual investigators might want to judge their particular acceptable levels of disparity.¹⁵ Another caveat centres on potential interim analyses done on sample sizes of less than 200—ie, before investigators reach total sample size. Greater

relative disparities in treatment group sizes could materialise in those instances, although we feel those costs are more than offset by the gains in unpredictability from simple randomisation.

For non-double-blinded randomised controlled trials with a sample size of less than 200 overall or within any principal stratum or subgroup of a stratified trial, we recommend a restricted randomisation procedure. The urn design⁴ functions especially well to promote balance without forcing it.¹⁶ It tends to balance more in the important early stages of a trial and then approach simple randomisation as the trial size increases. This attribute becomes useful with uncertain overall trial sizes, or more likely, uncertain stratum sizes in a stratified trial. It also proves useful in trials that might be ended due to sequential monitoring of treatment effects. Urn designs usually have adequate balancing properties while still being less susceptible to selection bias than permuted-block designs.^{4,14}

With these desirable properties come caveats. Some statisticians recommend use of permutation tests³ with urn randomisation designs. Permutation tests are assumption-free statistical tests of the equality of treatments.^{3,16} Unfortunately, they usually are not available for urn designs in standard statistical software.¹⁶ That adds analytical complexity for researchers and statisticians. However, if no major time trends on the outcome variables exist, use of standard statistical analyses from widely available software on trials that use urn randomisation would normally yield similar results to permutation tests.¹⁶ Moreover, with standard statistical analyses, investigators can easily obtain confidence intervals for common measures of effect.

Surprisingly, urn randomisation or other forms of biased-coin designs appear infrequently in reports.⁴ Perhaps another impediment to widespread usage pertains to the conceptual complexities of urn randomisation; it is more difficult to understand than simple or permuted-block randomisation. Whatever the reasons, urn designs languish in obscurity.

Mixed randomisation

Researchers should have an unpredictable approach to use in non-double-blinded trials until they feel comfortable with

urn randomisation or other approaches that enhance unpredictability.¹⁷ We have attempted to identify an approach that builds on the existing knowledge in clinical epidemiology. We propose a restricted randomisation method that should approach the unpredictability of urn randomisation overall while exceeding its unpredictability for small sample sizes, but without its real and perceived complexities. Our proposed approach promotes balance, but not the perfect balance that we feel can lead to predictability and selection bias.

Our solution mixes simple randomisation with permuted-block randomisation. Simple randomisation contributes the unpredictability to the approach whereas permuted-block randomisation contributes the balance. Our mixed approach begins with an uneven block generated by a replacement randomisation procedure (panels 2 and 3).¹⁸⁻²¹ Then, in its simplest form, standard permuted blocks of varying size follow. The replacement randomisation sequence would establish inequality initially and make any anticipation of assignments improbable throughout the remainder of the trial.

Replacement randomisation is basically simple randomisation, with a slight twist.⁴ Since investigators should aim for an uneven block, they would select a prespecified inequality in the sample sizes of the allocated groups. Then they would prepare an allocation sequence by simple randomisation, and check the disparity in sample sizes against their prespecified inequality. If the disparity in sample sizes meets or exceeds their prespecified disparity, then that simple randomisation allocation sequence suffices for the first uneven block. If not, then a whole new simple randomisation list is generated to replace the previous one. They would iterate until a simple randomisation sequence meets or exceeds their prespecified disparity (panel 2). The block size of the first uneven block could be odd or even overall and could be of most any total size, although we conceive of it usually falling in the range of five to 16.

Our basic approach creates an initial imbalance in a trial. Replacement randomisation represents just one approach to creating it. Another choice might be to select from random imbalanced permuted blocks or several variations on that

Panel 2: Mixed randomisation steps

Step 1: Generate one uneven block by replacement randomisation for the first participants

- A Identify block size for the first uneven block. The block size can be odd or even and of any reasonable size, but usually in the range of five to 16.
- B Select a prespecified inequality in the sample sizes of the allocated groups for the first uneven block.
- C Generate a simple randomisation sequence (eg, with a table of random numbers or a computer random number generator).
- D Inspect resultant sequence of assignments for matching or exceeding the desired prespecified inequality from Step B above.
- E If sufficiently unequal distribution of As and Bs, proceed to step 2; if not, go back to Step C above (iterate).

Step 2: Generate random permuted blocks for subsequent participants

- A Select block sizes for the permuted blocks. Longer block sizes, such as ten to 20, are more unpredictable than shorter block sizes, such as two to four. Longer block sizes should be preferred, unless an investigator needs approximate balance in a small trial or a small stratum of a trial. For example, an investigator might select block sizes of eight, ten, 12, and 14 as options.
- B Generate random permuted blocks, randomly varying the block size, as described in many texts.¹⁸⁻²¹
- C Decide if an additional uneven block or simple randomisation block is to be interspersed in the trial. If not, complete the required sample size with random permuted blocks. Otherwise, identify a point at which to interject an uneven block or a simple randomised sequence.
- D If interjecting another uneven block by replacement randomisation, proceed back to step 1. If interjecting a simple random sequence, proceed to step 3.

Step 3: Generate a simple random sequence for interjection after a set of permuted blocks

- A Identify the size of this simple random sequence. The size can be odd or even and of any reasonable size, but usually in the range of five to 16. We suggest an odd number to ensure some imbalance.
- B Generate a simple random sequence of the chosen size as suggested earlier.^{4,21}
- C Proceed to step 2, B.

Panel 3: Example of mixed randomisation

The randomisation scheme required an overall sample size of 100, based on power calculations. The investigator decided on a first uneven block size of ten and prespecified an inequality of at least four participants between treatment A and treatment B in that first uneven block. The investigator then proceeded to use replacement randomisation by successively selecting a simple random sequence⁴ of ten until that process yielded a sequence of ten assignments with either treatment A or treatment B having at least four extra participants. That sequence was B, A, B, B, B, A, B, B, A, B; actually, three treatment As and seven treatment Bs.²¹ Then the investigator decided to randomly vary the permuted-block sizes between six, eight, ten, and 12, as described in many sources.¹⁸⁻²¹ That could simply be continued over the remainder of the study, but this investigator decided to interject a simple randomised sequence of five assignments after a whole permuted-block passed the 40th participant. The block sizes randomly selected were, in order, 12, eight, and ten. The simple randomised sequence beginning with the 41st participant was B, A, B, A, A. After that simple randomised sequence, the investigator again proceeded with random permuted blocks of six, eight, ten, or 12, with the first randomly selected block size of eight. We depicted the allocation sequence and the total cumulative assignments by treatment for the first 53 assignments:

		Assignment number	Assignment	Cumulative				Assignment number	Assignment	Cumulative	
				A	B					A	B
Replacement randomisation block of 10		1	B	0	1			31	B	13	18
		2	A	1	1			32	A	14	18
		3	B	1	2			33	B	14	19
		4	B	1	3			34	B	14	20
		5	B	1	4			35	A	15	20
		6	A	2	4			36	B	15	21
		7	B	2	5			37	B	15	22
		8	B	2	6			38	A	16	22
		9	A	3	6			39	A	17	22
		10	B	3	7			40	A	18	22
Random permuted block of 12		11	A	4	7			41	B	18	23
		12	A	5	7			42	A	19	23
		13	B	5	8			43	B	19	24
		14	A	6	8			44	A	20	24
		15	A	7	8			45	A	21	24
		16	A	8	8			46	A	22	24
		17	A	9	8			47	A	23	24
		18	B	9	9			48	B	23	25
		19	B	9	10			49	A	24	25
		20	B	9	11			50	A	25	25
Random permuted block of 8		21	B	9	12			51	B	25	26
		22	B	9	13			52	B	25	27
		23	A	10	13			53	B	25	28
		24	B	10	14			54	-	-	-
		25	A	11	14			55	-	-	-
		26	A	12	14			56	-	-	-
		27	B	12	15			57	-	-	-
		28	B	12	16			58	-	-	-
		29	A	13	16			59	-	-	-
		30	B	13	17			60	-	-	-

Assignment numbers, sequence assignments, and cumulative assignments for the mixed randomisation example, with the successive blocks of assignments alternately bolded.

theme (Douglas G Altman, personal communication). Another excellent easy approach, if investigators can accept variation in the disparity, would involve setting the overall first block size at an odd number, which ensures at least some disparity, and just using simple randomisation (without replacement randomisation).

Identification of an acceptable prespecified inequality for the first block is quite insensitive. Power remains robust up to about a two-to-one ratio for the total sample sizes in the treatment groups.³ Investigators need only create much smaller inequalities than that, particularly in small strata in a stratified trial. Ensuring unpredictability will probably happen as much from small inequalities as large inequalities. Moreover, inequalities can actually slightly increase power in addition to offering enhanced unpredictability. For example, in tests for proportions or life-tables, the maximum power is attained with unequal treatment group sizes.³

After the first uneven block, investigators should proceed as in normal permuted-block randomisation (panels 2 and 3). We suggest that they randomly vary the block size and use as long a block size as practicable for greater unpredictability.⁴ For added unpredictability, investigators could also intersperse additional uneven blocks generated with replacement randomisation throughout the trial. For example, another uneven block could be interjected after a permuted-block surpasses the next 50 participants (using the whole block, which means that the next uneven block would likely begin beyond 50). Alternatively, for these interspersed blocks, investigators could just use simple randomisation. That would be slightly easier, likely provide additional unpredictability, and also provide a richer set of potential allocation sequences. Other potential options for these interspersed uneven blocks exist—eg, imbalanced permuted blocks—but they extend beyond the range of this essay.

For analysis, we suggest use of standard statistical analyses with readily available statistical software—ie, the simpler approach. Slightly greater credibility for hypothesis testing might be gained with design-based permutation tests,¹¹ but we favour the estimates with confidence intervals that standard statistical analyses produce.¹⁴ We also agree with the acceptability of ignoring blocking in the analysis.¹⁸ This straightforward approach usually produces slightly conservative results in trials that use blocking, if a time trend on the outcome exists, but otherwise produces similar results to an analysis that incorporates blocking.¹¹

Those issues, however, pale in comparison to the potential effects of selection bias that could arise in the absence of unpredictability. Once selection bias infiltrates a trial, it becomes imbedded and usually prevails undetected, except in limited situations where investigators might use an innovative detection approach.²² Moreover, empirical evidence points toward substantial selection biases.^{5,6,23} However, the discussions of standard statistical analyses versus permutation tests, or unblocked analyses versus blocked analyses from trials with permuted-block randomisation focus on seemingly smaller increments of p values or power. Irrespective of the analysis method chosen, the interpretation from a trial in many instances would be the same. More importantly, unlike selection bias, investigators have a straightforward remedy. If a journal editor or statistical reviewer insists on a different approach, investigators can usually retreat to a blocked analysis or a permutation test. In sum, investigators should use their energy to focus on prevention of biases in the design and implementation of their trial, with an unpredictable allocation sequence being an integral part of that effort.

Full disclosure in the protocol?

Provision of explicit details of the randomisation scheme in the protocol might facilitate deciphering of the allocation sequence, thus undermining the process. We recommend that researchers not fully describe their generation scheme in their research protocol and investigators manual. They would have to describe any stratification plans, but those implementing the trial should be kept ignorant of the full details of the method to generate the allocation sequence.

Some funding authorities require more documentation to ensure that researchers know proper randomisation methods. Appropriate rationale and references might suffice. If the funding agency requires more specifics, a researcher should provide a separate generation of the allocation sequence plan to the funders that will not be shared with those enrolling participants. However, in the final trial report, researchers should fully document the randomisation approach.^{13,24}

Conclusion

Investigators underuse simple randomisation and overuse fixed-blocked randomisation. They do so because they inadequately appreciate the importance of unpredictability and overvalue equal treatment group sizes. Simple randomisation is totally unpredictable, implements easily, and enables use of standard statistical analysis software. For non-double-blinded trials larger than 200 participants, investigators should use it more and tolerate, if not celebrate, the disparity in group sizes. Such unpredictability reflects the essence of randomness.

For non-double-blinded randomised controlled trials smaller than about 200 participants overall or within any principal stratum or subgroup, the urn design enhances unpredictability compared with blocking. Our mixed randomisation method, however, attains unpredictability within the context of the currently understood simple and

permuted-block randomisation methods. We urge researchers to use our method, at least in non-double-blinded trials.

Why add complexity to implementation of trials? The answer resides in the overriding importance of protecting the integrity of randomisation. Proper randomisation minimises bias, more than any other methodological aspect of a trial: “When the randomization leaks, the trial’s guarantee of lack of bias runs down the drain.”²⁵ Those involved in trials go to great pains to decipher randomisation schemes.^{5,6,23} Accordingly, researchers who design trials must take equally great pains to thwart those efforts.

We thank Willard Cates, David L Sackett, Douglas G Altman, Rosalie Dominick, and Vance W Berger for their helpful comments on an earlier version of this report. Much of the material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

- Keirse MJNC. Electronic monitoring: who needs a trojan horse? *Birth* 1994; **21**: 111–13.
- Cates W Jr, Grimes DA, Schulz KF, Ory HW, Tyler CW Jr. World Health Organization studies of prostaglandins versus saline as abortifacients: a reappraisal. *Obstet Gynecol* 1978; **52**: 493–98.
- Lachin JM. Statistical properties of randomization in clinical trials. *Control Clin Trials* 1988; **9**: 289–311.
- Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. *Lancet* 2002; **359**: 515–19.
- Schulz KF. Subverting randomization in controlled trials. *JAMA* 1995; **274**: 1456–58.
- Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273**: 408–12.
- Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. *Lancet* 2002; **359**: 614–18.
- Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA* 1994; **272**: 125–28.
- Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. *Lancet* 1990; **335**: 149–53.
- Adetugbo K, Williams H. How well are randomized controlled trials reported in the dermatology literature? *Arch Dermatol* 2000; **136**: 381–85.
- Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. *Control Clin Trials* 1988; **9**: 327–44.
- Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient, I: introduction and design. *Br J Cancer* 1976; **34**: 585–612.
- Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; **134**: 663–94.
- Lachin JM, Matts JP, Wei LJ. Randomization in clinical trials: conclusions and recommendations. *Control Clin Trials* 1988; **9**: 365–74.
- Lachin JM. Properties of simple randomization in clinical trials. *Control Clin Trials* 1988; **9**: 312–26.
- Wei LJ, Lachin JM. Properties of the urn randomization in clinical trials. *Control Clin Trials* 1988; **9**: 345–64.
- Berger VW, Ivanova A, Wei EY, Knoll MAD. An alternative to the complete randomized block procedure. *Control Clin Trials* 2001; **22**: 43S.
- Friedman L, Furberg C, DeMets D. Fundamentals of clinical trials. St Louis: Mosby, 1996.
- Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley, 1983.
- Meinert CL. Clinical trials: design, conduct, and analysis. New York: Oxford University Press, 1986.
- Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991.
- Berger VW, Exner DV. Detecting selection bias in randomized clinical trials. *Control Clin Trials* 1999; **20**: 319–27.
- Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998; **352**: 609–13.
- Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports or parallel-group randomised trials. *Lancet* 2001; **357**: 1191–94.
- Mosteller F, Gilbert JP, McPeck B. Reporting standards and research strategies for controlled trials: agenda for the editor. *Control Clin Trials* 1980; **1**: 37–58.