

accelerated factor See SURVIVAL ANALYSIS

accelerated failure time models See SURVIVAL ANALYSIS, TRANSFORMATION

active control equivalence studies The classic randomised CLINICAL TRIAL seeks to prove superiority of a new treatment to an existing one and a successful conclusion is one in which such proof is demonstrated. The famous MRC trial of streptomycin is a case in point (Medical Research Council Streptomycin in Tuberculosis Trials Committee, 1948). The trial concluded with a significant difference in outcome in favour of the group given streptomycin compared to the group that was not. In recent years, however, there has been an increasing interest in trials whose objective is to show that some new therapy is no worse as regards some outcome than an existing treatment. Such trials have particular features and difficulties that were described in an important paper by Makuch and Johnson (1989) in which they used the term 'active control equivalence studies' (ACES).

Actually, the term is not ideally chosen since, unlike bioequivalence studies, where the object is to show that the bioavailability of a new formulation is not only at least 20% less than that of an existing formulation, but also at most 25% more, and hence where *equivalence* to some degree is genuinely the aim, in ACES it is almost always the case that only noninferiority is the goal. It may be questioned as to why the rather modest goal of noninferiority should be of any interest in drug regulation. There are several reasons. The first is that the new drug may have advantages in terms of tolerability. Second, the new drug, while showing no net advantage to the existing one, may increase patient choice and this can be useful. For example, many people have an aspirin allergy. Hence, it is desirable to have alternative analgesics, even if no better on average than aspirin. Third, it may become necessary to withdraw treatments from the market and one can never predict when this may happen. There are now several statins on the market. The fact that this is so means that withdrawal of cerivastatin does not make it impossible for physicians to continue to treat their patients with this class of drug. Fourth, introduction of further equivalent therapies before patent expiry of an innovator in the class may permit price competition to the advantage of reimbursors (although such competition is probably not particularly effective; Senn and Rosati, 2003). However, the fifth reason is probably the most important. Drug regulation is designed to satisfy some minimum requirements for phar-

maceuticals: that they are of sufficient quality, are safe and efficacious. Efficacy is demonstrated if the treatment is better than placebo, even if it is not as good as some other treatments. The comparison of a new drug to an active treatment may be dictated by ethics but the object of the trial may simply be an indirect proof that the treatment is better than placebo through comparison to an agent whose efficacy is accepted.

Recently the issue of the indirect comparison to placebo has been taken more seriously. Consider the case where we have a single effective treatment on the market, say A, whose efficacy has been demonstrated in a series of trials comparing it to placebo. We now run some new trials comparing a further treatment, B, to A. Taking all these trials together, they then have the structure of an incomplete blocks design. The effect of B compared to placebo can then be estimated using the double contrast of B compared to A and A compared to placebo. This approach has been examined in detail by Hasselblad and Kong (2001). A consequence of taking this particular view of matters is that the precision with which the effect of A was established compared to placebo cannot be exceeded by the indirect comparison of B to placebo, since the variance of this indirect contrast is the sum of the variances of the two direct contrasts.

This is, however, not the only difficulty with such studies. The following are some of those that apply.

Establishing a clinically irrelevant difference. If the route of a formal analysis compared to placebo via an indirect contrast is taken, this particular difficulty may be finessed. The new treatment is shown to be 'significantly' better than placebo, albeit using an indirect argument, and the extent of its inferiority to the comparator is only of relevance to the extent that it impinges on the proof of efficacy compared to placebo. If this proof is provided, then the comparison to the active comparator is 'water under the bridge'. If this particular approach is *not* taken, however, then any proof of efficacy of the new treatment rests on a demonstration that it is not 'substantially inferior' to the comparator, which comparator is accepted as being efficacious. This raises the issue as to what it means for a drug to be not substantially inferior to another one. This appears to require that some margin Δ , $\Delta > 0$, be adopted such that if τ is the extent by which the new treatment is inferior to the standard (where $\tau < 0$ indicates inferiority) then it is judged *substantially* inferior if $\tau \leq -\Delta$ and not substantially inferior or 'equivalent' if $\tau > -\Delta$.

Technical statistical aspects. In a Neyman–Pearson framework (see Salsbury, 1998) the test of noninferiority requires one to use a shifted NULL HYPOTHESIS. One might, therefore, adopt $H_0: \tau \leq -\Delta$. The situation is not as controversial as that for true bioequivalence, where the fact that two hypotheses have to be rejected, that of inferiority and that of superiority, means that an intuitive approach of seeing that the confidence limits for the difference lie within the limits of equivalence is not ‘optimal’ (Berger and Hsu, 1996), although the ‘optimal’ test may in practice be worse (Perlman and Wu, 1999; Senn, 2001). In practice, in the case of ACES if the lower conventional $1 - \alpha$ two-sided CONFIDENCE INTERVAL for τ exceeds $-\Delta$, the hypothesis of substantial inferiority may be rejected at the level α and noninferiority asserted. It might be thought that a one-sided confidence interval would be sufficient for this purpose. However, the general regulatory convention is that all tests designed to show superiority are two-sided (despite apparent purpose) and, since such tests are a special case of a noninferiority test with $\Delta = 0$, use of one-sided tests for noninferiority would lead to inconsistencies (Senn, 1997; Committee for Proprietary Medicinal Products, 2000). In a Bayesian framework (see BAYESIAN METHODS) one might require that the posterior probability of noninferiority were less than some specified amount. Alternatively, use of a loss function would permit a decision analytic method, such as has been proposed for bioequivalence (Lindley, 1998), to be used.

Power of trials. Note that the reason one does not employ a value of $\Delta = 0$ in practice is that unless it is expected that the new treatment really is better than the standard, the power of the resulting test could never exceed 50%. However, the clinically irrelevant difference is likely to be less than the clinically relevant difference used in conventional trials. Hence, if the new treatment is actually no better than the standard treatment, then, for a given sample size, the noncentrality parameter, $\delta = \Delta / SE(\hat{\Delta})$ is likely to be smaller for ACES than for trials designed to show superiority. Consequently, ACES either have lower power or higher sample sizes than conventional trials.

Assay sensitivity. A problem with ACES is that if the trial appears to show noninferiority of the new treatment, then there are three plausible explanations. The first, that of chance, is one that statistical analysis is designed to address. The second, that the new treatment is indeed noninferior, is what was desired to prove. However, a third possibility, that the experiment was not sensitive to find a difference, is difficult to exclude. This issue has been referred to as one of ‘competence’ (Senn, 1993) and affects whatever

inferential framework one decides to use. An analogy may be useful here. In a game of hunt the thimble, a found thimble renders the quality of the strategy used for finding it irrelevant. It is no more ‘found’ if a good strategy were used than if a bad one were. However, a failure to find a thimble does not automatically justify the conclusion that the room does not contain one and the quality of the search employed is a crucial consideration in any judgement that it does not.

The effect of DROPOUTS, NONCOMPLIANCE and the role of INTENTION-TO-TREAT analysis. It is plausible that in many circumstances in conventional superiority trials if noncompliance or dropouts are a problem an intention-to-treat analysis will give a more modest estimate of the treatment effect than will a PER PROTOCOL analysis. In ACES, it is at least plausible that this may not be the case.

Conflict of requirements of additivity and clinical relevance. It may be that the clinically irrelevant difference is most meaningfully established on a scale that is not additive. For example, in a trial of an anti-infective, it could be most appropriate to establish that the difference in cure rate on the probability scale was not greater than some specified amount. Contrariwise, the log-odds scale might lend itself more readily to statistical modelling. This can lead to considerable difficulties (Holmgren, 1999), in particular because a trial does not recruit a random sample from the target population. It may be that further modelling using additional data may be necessary (Senn, 2000).

A common circumstance likely to make regulatory authorities ask questions is that a trial that was designed with optimism to show superiority to an active comparator fails to do so, but then is used to attempt to demonstrate noninferiority. This particular set of circumstances has become the subject of one of the European Medicine Evaluation Agency’s ‘points to consider’ (Senn, 1997; Committee for Proprietary Medicinal Products, 2000). This stresses the desirability of establishing the trial’s purpose pre-performance and also warns against establishing the clinically irrelevant difference, Δ , after the trial is complete. It regards putting a trial that was designed to show superiority to the purpose of noninferiority as an unacceptable use but accepts the converse. The guideline recognises that there are no issues of multiple testing involved with such switches (Bauer and Kieser, 1996) but that establishing values of Δ retrospectively may be biasing. Thus, it is preferable for investigators to specify in advance (e.g. by means of formal change to the CLINICAL TRIALS PROTOCOL) their intended switch of purpose and to fix the value of Δ prior to data unblinding. This, however, raises the issue as to whether

the value of Δ is not something the regulator should declare for given indications rather than relying on the sponsor to do so. Otherwise, a regulator could be faced with the following position. Drug B is registered on the basis of comparison to a standard treatment A because the lower confidence interval for the treatment effect, τ_{B-A} , exceeds some pre-specified value Δ . However, a further drug, C, which has also been compared to A, is not granted a licence because a superiority trial was planned. Although superiority to A was not proven, the lower confidence interval for the treatment effect τ_{C-A} excludes a smaller possible difference between C and A than is excluded for the difference between B and A by the trial that has led to registration of B. SS

Bauer, P. and Kieser, M. 1996: A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika* 83, 4, 934–7. **Berger, R. L. and Hsu, J. C.** 1996: Bioequivalence trials, intersection–union tests and equivalence confidence sets. *Statistical Science* 11, 4, 283–302. **Committee for Proprietary Medicinal Products** 2000: Points to consider on switching between superiority and non-inferiority. **Hasselblad, V. and Kong, D. F.** 2001: Statistical methods for comparison to placebo in active-control studies. *Drug Information Journal* 35, 435–49. **Holmgren, E. B.** 1999: Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. *Journal of Biopharmaceutical Statistics* 9, 4, 651–9. **Lindley, D. V.** 1998: Decision analysis and bioequivalence trials. *Statistical Science* 13, 2, 136–41. **Makuch, R. and Johnson, M.** 1989: Issues in planning and interpreting active control equivalence studies. *Journal of Clinical Epidemiology* 42, 6, 503–11. **Medical Research Council Streptomycin in Tuberculosis Trials Committee** 1948: Streptomycin treatment for pulmonary tuberculosis. *British Medical Journal* ii, 769–82. **Perلمان, M. D. and Wu, L.** 1999: The emperor’s new tests. *Statistical Science* 14, 4, 355–69. **Salsbury, D.** 1998: Hypothesis testing. In Armitage, P. and Colton, T. (eds), *Encyclopedia of biostatistics*. Chichester: John Wiley & Sons, Ltd. **Senn, S. J.** 1993: Inherent difficulties with active control equivalence studies. *Statistics in Medicine* 12, 24, 2367–75. **Senn, S. J.** 1997: *Statistical issues in drug development*. Chichester: John Wiley & Sons, Ltd. **Senn, S. J.** 2000: Consensus and controversy in pharmaceutical statistics (with discussion). *The Statistician* 49, 135–76. **Senn, S. J.** 2001: Statistical issues in bioequivalence. *Statistics in Medicine* 20, 17–18, 2785–99. **Senn, S. J. and Rosati, N.** 2003: Editorial: Pharmaceuticals, patents and competition – some statistical issues. *Journal of the Royal Statistical Society Series A – Statistics in Society* 166, 271–7.

adaptive designs CLINICAL TRIALS that are adaptive are modified in some way by the data that have already been collected within that trial. The most common way the designs adapt is in the allocation of treatment, as a function of the response. For example, we may be interested in a dose that gives a 20 % chance of toxicity, where excesses to this level of toxicity would be harmful. Therefore, we may want to design the trial in such a way that, as more information is gathered, doses are allocated to optimise the estimate of that dose. If we

were to use a traditional fully randomised approach to running the trial, which is not adaptive, we would probably not look at the data until the end of the trial, thereby risking exposing subjects to toxic doses and also possibly failing to produce an optimal estimate of the required dose. Another such example of an adaptive design is given in Rosenberger and Lachin (1993), whereby there are two treatments in the study, A and B, and as information emerges from the trial the treatment assignment probabilities are adapted in an attempt to assign more patients to the treatment performing better thus far. Therefore, when a patient enters the study, if treatment A appears to be better than treatment B, a patient has a greater than 50 % chance of being allocated treatment A – and vice versa.

Because adaptive designs modify the allocation of treatment on an ongoing basis, and thus protect patients from ineffective or toxic doses, they can be said to be more ethical than traditional designs. Rosenberger and Palmer (1999) consider the ethical dilemma between collective and individual ethics (see ETHICS AND CLINICAL TRIALS) and argue that in a clinical trial setting individual ethics should be uppermost; i.e. consideration should be towards doing what is best for patients in the current trial as opposed to doing what is best for future patients who stand to benefit from the results of current trial. The Declaration of Helsinki of October 2000 outlines the tension between these two types of ethics by stating: ‘Considerations related to the well-being of the human subject should take precedence over the interests of science and society.’ It is adaptive designs that address the individual ethics, as opposed to fully randomised designs, which address those collective ethics.

We will be dealing primarily with response adaptive designs here, such as those just outlined, and will not be describing those designs that attempt dynamically to balance the randomisation for covariate information, such as outlined by Pocock and Simon (1975) (see DATA-DEPENDENT DESIGNS, MINIMISATION).

The randomised play winner (RPW) design attempts to allocate treatments to patients sequentially based on a simple probability model. Rosenberger (1999) emphasises that the RPW design specifically applies to the situation where the outcome from a trial is binary, i.e. either ‘success’ or ‘failure’ and where there are only two treatments, e.g. drug A and drug B. At the start of the trial there is an assumed urn of α balls of type A (which relate to drug A) and β balls of type B (which relate to drug B). When a subject is recruited, a ball is drawn from the urn and then replaced. If the ball is type A then the subject is allocated to drug A, if type B then the subject is allocated to drug B. When the subject’s outcome is available (and we assume that the outcome is available before the next subject is randomised), the urn is updated. If the response is a success on drug A, then a ball of type A is put into the urn, and

similarly for a success on drug B. If the outcome is a failure on drug A, then a ball of type B is put into the urn, and again similarly for a failure on drug B. In this way, the balls build up such that a new subject has a better chance of being allocated to a better treatment.

Rosenberger (1999) concludes with a table of conditions under which the RPW rule is reasonable and provides a realistic alternative to the standard clinical trial design. These are given in the table.

adaptive designs *Conditions under which the RPW is reasonable (Rosenberger, 1999)*

- The therapies have been evaluated previously for toxicity
- The response is binary
- Delay in response is moderate, allowing adapting to take place
- Sample sizes are moderate (at least 50 subjects)
- Duration of the trial is limited and recruitment can take place during the entire trial
- The trial is carefully planned with extensive computations done under different models and initial urn compositions
- The experimental therapy is expected to have significant benefits to public health if it proves effective

Traditional dose–response studies, where patients are allocated to a limited number of doses along an assumed dose–response curve, are limited and, some would say, wrong. For example, if the assumed dose–response model is incorrect then patients may be allocated to ineffective or unsafe doses. One answer could be to increase the number of doses. However, this would result in many patients allocated to wasted doses. It would be much better to increase the number of doses and allocate doses to a subject based on current knowledge of the dose–response curve, which best optimises some pre-specified criteria. This is precisely what Bayesian response adaptive designs attempt to do, by employing Bayesian DECISION THEORY to a utility function. Thus, the dose that most optimally addresses the utility is allocated to the next available subject or cohort of subjects.

One of the first BAYESIAN METHODS described was the continual reassessment method (CRM), introduced by O’Quigley, Pepe and Fisher (1990), and originally devised for dose-escalation studies in oncology. Whitehead *et al.* (2001a) suggest that the method could also be used for applications in other serious diseases. The CRM envisages a study whereby human volunteers are treated sequentially, in order to detect a dose with a probability of toxicity of 20%, i.e. TD20. The response is a binary response, ‘toxicity’ or ‘no toxicity’. Before the study starts, investigators are asked to provide what their best guess is of a probability of toxicity at

each of the series of doses. The first patient is then treated with the dose that is considered to be the closest to the TD20. Once the outcome is observed the PROBABILITY of toxicity at each of the doses is recalculated using the Bayesian method of statistics. The procedure continues in this way until it settles on a single dose. Whitehead *et al.* (2001a) point out that the CRM does home in on the TD20 quickly and efficiently, but there has been concern that early on in the trial subjects could be allocated to too high a dose, leading to potential toxicity problems. This has led to a number of modifications, such as starting at the lowest dose and never skipping a dose during the escalation.

Whitehead *et al.* (2001b) suggest practical extensions to the CRM for pharmacokinetic data, employing the use of Bayesian decision theory to allocate treatments optimally to subjects. They argue that conventional dose-escalation studies carried out in healthy volunteers do not normally employ statistical methodology or formal guidelines for dose escalation. As such the studies can take a long time to complete with little opportunity to skip doses. The methods proposed allocate doses in order to maximise the information about the dose–response curve, given a pre-specified safety constraint. They use two simple utility or gain functions, one that allocates the highest allowable dose under the safety constraint and the other that allocates doses in order to optimise the shape of the dose–response curve.

Krams *et al.* (2003) also use a Bayesian decision theory approach with sequential dose allocation to a Phase II study in acute stroke therapy by inhibition of neutrophils (ASTIN), which employs up to 15 dose levels. They use a response-adaptive procedure in order to find a dose that gives an improvement over that of placebo in the primary ENDPOINT, allocating the next subject either to the optimal dose or PLACEBO. Stopping rules were employed by which if the posterior probability of an effective drug or ineffective drug were greater than 0.9 then the decision would be made either to go on to a confirmatory trial (effective drug) or to stop development (ineffective drug). In this way, they were able to stop development of a compound more quickly than would have been possible under the traditional paradigm.

In 2006, the Pharmaceutical Research and Manufacturers of America (PhRMA) Adaptive Design Working Group published a series of papers in an issue of the Drug Information Association (DIA) journal detailing various aspects of these trials. Topics included terminology and classification; implementation; confidentiality and trial integrity; adaptive dose response; seamless Phase II/III; and sample size reestimation (see *Drug Information Journal* 40, 425–84, 2006). In addition, and reflecting the growing interest in adaptive designs, there have been numerous special editions of other journals devoted to these trials, including *Journal of Statistical Planning and Inference*, issue 136(2), 2006; *Journal of Biopharmaceutical Statistics*, issues 16(5),

2006 and 17(6), 2007; and *Statistics in Medicine*, issue 27(10), 2008. AB

Krams, M., Lees, K., Hacke, W., Grieve, A. P., Orgogozo, J.-M. and Ford, G. A. 2003: Acute stroke therapy by inhibition of neutrophils (ASTIN). An adaptive dose-response study of UK-279,276 in acute ischemic stroke. *Stroke* 34, 2543–8. **Pocock, S. and Simon, R.** 1975: Sequential treatment assignment with balancing of prognostic factors in controlled clinical trials. *Biometrics* 31, 103–15. **O’Quigley, J., Pepe, M. and Fisher, L.** 1990: Continual reassessment method: a practical design for Phase I clinical trials in cancer. *Biometrics* 46, 33–48. **Rosenberger, W. F.** 1999: Randomized play-the-winner clinical trials: review and recommendations. *Controlled Clinical Trials* 20, 328–42. **Rosenberger, W. F. and Lachin, J. M.** 1993: The use of response-adaptive designs in clinical trials. *Controlled Clinical Trials* 14, 471–84. **Rosenberger, W. F. and Palmer, C. R.** 1999: Ethics and practice: alternative designs for Phase III randomised clinical trials. *Controlled Clinical Trials* 20, 172–86. **Whitehead, J., Yinghui, Z., Patterson, S., Webber, D. and Francis, S.** 2001a: Easy-to-implement Bayesian methods for dose-escalation studies in healthy volunteers. *Biostatistics* 2, 47–61. **Whitehead, J., Zhou, Y., Stallard, N., Todd, S. and Whitehead A.** 2001b: Learning from previous responses in Phase I dose-escalation studies. *British Journal of Clinical Pharmacology* 52, 1–7.

adaptive randomisation See ADAPTIVE DESIGNS, RANDOMISATION

adjustment for noncompliance in randomised controlled trials In clinical medicine, ‘noncompliance’ occurs when a patient does not fully follow a prescribed course of treatment. The alternative terms ‘adherence’ and ‘concordance’ attempt to avoid the authoritarian overtones of ‘compliance’. In randomised CLINICAL TRIALS, we are concerned with any departure from a randomised treatment, whether due to noncompliance or a treatment change agreed with medical staff. In a trial to compare two types of medication (drug A and drug B, say) for the treatment of heart disease, for example, patients may refuse or forget to take any of their medication or forget to take it some of the time (partial compliance). Patients allocated to receive drug A might switch to drug B, and vice versa. Some of the patients may even take another medication altogether (drug C, say) or, particularly if the therapy appears to be failing, receive a much more radical intervention such as surgery. A further complication for the estimation of treatment effects arises when patients who fail to comply with their prescribed treatment are also those who are more likely to be lost to follow-up.

Rationale. Conventionally, trials with departures from randomised treatment are analysed by INTENTION-TO-TREAT. This directly compares the *effectiveness* of the different

treatment policies as actually implemented in the trial – e.g. ‘drug A plus changes’ versus ‘drug B plus changes’. Unlike effectiveness, *efficacy* relates to the effects of the treatments themselves, and is not estimated by an intention-to-treat analysis. Researchers may also be interested in the effectiveness of an intervention in other circumstances, e.g. if public suspicion of the intervention had been reduced by the positive results of a clinical trial. In these circumstances, the rates of compliance may be improved and adjustment for this change may be attempted.

It is important to define the aim of adjustment for non-compliance. For example, in a trial of immediate versus deferred zidovudine in asymptomatic HIV infection, the initial plan was to defer zidovudine until the onset of symptomatic disease. However, following a protocol amendment, some individuals started zidovudine before the onset of symptomatic disease (White *et al.*, 1997). There was interest in estimating the effect that would have been observed under the original protocol. Zidovudine before the onset of symptomatic disease was therefore regarded as ‘noncompliance’. Other individuals stopped zidovudine treatment because of adverse events. Additional adjustment for stopping treatment would not answer a clinically relevant question, so the analysis did not aim to estimate efficacy.

Adjustment for noncompliance is useful in a variety of situations. Patients may be most interested in treatment efficacy. Differences in compliance may help to explain variation of a treatment effect with time, between subgroups in a trial or between trials in a META-ANALYSIS. Reconciling trial data with observational data may require adjustment for noncompliance in the trial. Policy analysis may require projections for situations with improved compliance.

Most attempts to allow for noncompliance use on-treatment analysis or PER PROTOCOL analysis. This only provides a valid comparison of the treatments themselves (efficacy) if compliers and noncompliers do not differ systematically in their disease state or prognosis. In practice this is unlikely to be the case, so selection bias occurs. Heart disease patients who comply with their prescribed medication, for example, are also those who are likely to improve their diet or take more exercise and these changes, in turn, are likely to lead to a better outcome. SELECTION BIAS may often be reduced by adjustment for baseline covariates, but there is still no guarantee of an unbiased analysis. For example, in the Coronary Drug Project, 5-year mortality of poor compliers was 28.2% compared with 15.1% in good compliers, and adjustment for 40 baseline factors only reduced the difference to 25.8% versus 16.4% (The Coronary Drug Project Research Group, 1980).

Newer ‘randomisation-based’ methods can estimate efficacy while avoiding selection bias by directly comparing the groups as randomised as in an intention-to-treat analysis (White, 2005). This is made possible by considering the

subgroup of ‘compliers’ who would have received their randomised treatment, whichever group they were randomised to. For example, a trial in Indonesian children compared vitamin A supplementation with no intervention, the outcome being 12-month mortality. Vitamin A supplementation was actually received by only 80% of the intervention arm and by none of the control arm. Sommer and Zeger (1991) considered the subgroup who did not receive vitamin A in the intervention arm and a corresponding subgroup of the control arm who *would not have received vitamin A if they had been allocated to receive it*. These ‘noncomplier’ subgroups were assumed to be unaffected by allocation to vitamin A. It is then straightforward to estimate the number of noncompliers in the control arm and their mean outcome, and hence the risk difference, risk ratio or odds ratio in compliers. This is often called the ‘complier average causal effect’ (CACE) estimate (Little and Rubin, 2000). This approach is a special case of PRINCIPAL STRATIFICATION.

A more general approach requires a model relating *potential outcomes* for each individual under different counterfactual treatments. A simple model might say that each individual would have blood pressure b mmHg lower if they took the drug with perfect compliance than if they did not take the drug, with proportional blood pressure reductions for partial compliance. Such a model may be fitted by observing that untreated blood pressure must have the same distribution in each randomised group (Fischer-Lapp and Goetghebeur, 1999). An important advantage of these methods is that no assumption is required about the relationship between compliance and potential outcomes. They are closely related to the use of INSTRUMENTAL VARIABLES methods (Dunn and Bentall, 2007).

The approaches just described are generally only able to estimate one treatment effect in a two-arm trial. They tend to be hopelessly imprecise in situations such as EQUIVALENCE STUDIES where patients may stop all treatment during the trial, so that the analysis requires estimation of the effect of both treatments. In this case it is possible to adjust the randomised comparison using observational estimation of one or more treatment effects – i.e. assuming there are no unmeasured confounders for treatment. Methods such as *marginal structural modelling* can work even when actual treatment is both a consequence of symptomatic deterioration and a cause of slower disease progression (see Little and Rubin, 2000, for references to this literature).

A trial with noncompliance has less POWER than one with perfect compliance, as a result of the reduced effect size as estimated in an intention-to-treat analysis, and it is natural to want to recover the lost power. However, many of the new procedures preserve the intention-to-treat SIGNIFICANCE LEVEL and therefore do not affect power. In some cases, it is impossible to regain power without making some assumption

about comparability of noncompliers and compliers. In other situations, some gain in power is theoretically possible, but this is unlikely to be appreciable in practice (Becque and White, 2008). Significance testing should therefore rely on intention-to-treat analysis even when other methods are used to estimate efficacy. IW/GD

Angrist, J. D., Imbens, G. W. and Rubin, D. B. 1996: Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* 91, 444–72. **Becque, T. and White, I. R.** 2008. Regaining power lost by non-compliance via full probability modelling. *Statistics in Medicine* 27, 5640–63. **Dunn, G. and Bentall, R.** 2007: Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statistics in Medicine* 26, 4719–45. **Fisher-Lapp, K. and Goetghebeur, E.** 1999: Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Controlled Clinical Trials* 20, 531–46. **Little, R. and Rubin, D. B.** 2000: Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health* 21, 121–45. **The Coronary Drug Project Research Group** 1980: Influence of adherence to treatment and response to cholesterol on mortality in the Coronary Drug Project. *New England Journal of Medicine* 303, 1038–41. **White, I. R.** 2005: Uses and limitations of randomization-based efficacy estimators. *Statistical Methods in Medical Research* 14, 327–47. **White, I. R., Walker, S., Babiker, A. G. and Darbyshire, J. H.** 1997: Impact of treatment changes on the interpretation of the Concorde trial. *AIDS* 11, 999–1006.

age-period cohort analysis To understand the effect of time on a particular outcome for an individual it is essential to realise the relevant temporal perspective. Age affects many aspects of life, including the risk of disease, so this is an essential component of any analysis of time trends. Period denotes the date of the outcome and if the outcome varies with period it is likely to be due to some underlying factor that affects the outcome and varies in the same way for the entire population under study. Cohort, contrariwise, refers to generational effects caused by factors that only affect particular age groups when their level changes with time.

An example of a period effect would be a potential effect of an air contaminant that affected all age groups in the same way. If the level of exposure to that factor increased/decreased with time, exerting a change in the outcome in all age groups, then we would expect a related pattern across all age groups in the study. In studies that take place over long periods of time, the technology for measuring the outcome may change, giving rise to an artifactual effect that was not due to change in exposure to a causative agent. For example, intensive screening for disease can identify disease cases that would not previously have been identified, thus artificially increasing the disease rate in a population that has had no change in exposure over time.

Cohort (also called birth cohort) effects may be due to factors related to exposures associated with the date of birth, such as the introduction of a particular drug or practice during pregnancy that was brought in at a particular point in time. For example, a pregnancy practice associated with increased risk and adopted by the population of mothers during a particular time period could affect the risk during the lifespan of the entire generation born during that period. While it is common to refer to these effects as being associated with year of birth, they could also be the result of changes in exposure that occurred after birth. In many individuals, lifestyle factors that may affect disease risk over a lifetime are fixed as they approach adulthood. A quantification of these effects on such a generation would give rise to a comparison of these cohort or generational effects.

An inherent redundancy among these three temporal factors arises from the fact that knowing any two factors implies the value of the third. For example, if we know an individual's age (a) at a given date or period (p), then the cohort is the difference ($c = p - a$). This linear dependence gives rise to an identifiability problem in a formal regression model that attempts to obtain quantitative estimates of regression parameters associated with each temporal element:

$$E[Y] = \beta_0 + a\beta_a + p\beta_p + c\beta_c$$

Using the linear relationship between the temporal factors gives rise to:

$$\begin{aligned} E[Y] &= \beta_0 + a\beta_a + p\beta_p + (p-a)\beta_c \\ &= \beta_0 + a(\beta_a - \beta_c) + p(\beta_p + \beta_c) \end{aligned}$$

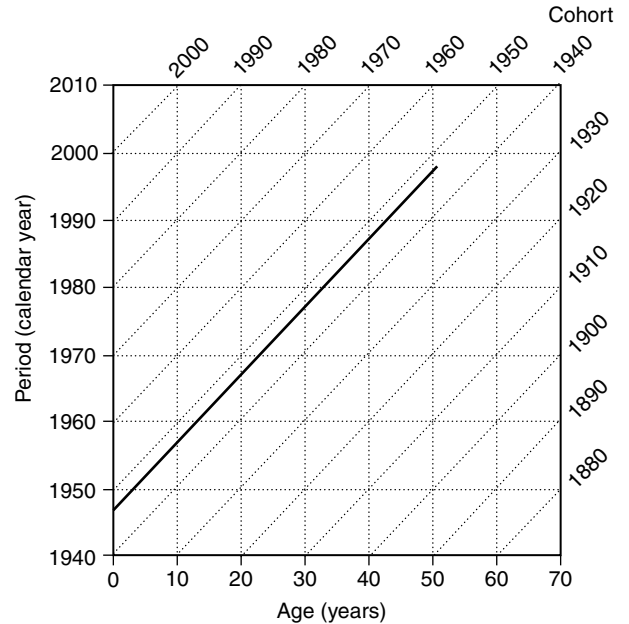
which has only two identifiable parameters besides the intercept instead of the expected three. Another way of visualising this phenomenon is that all combinations of age, period and cohort may be displayed in the LEXIS DIAGRAM (see the figure), which is obviously a representation of a two-dimensional plane instead of the three dimensions expected for three separate factors.

In general, these analyses are not limited to linear effects applied to a continuous measure of time, but instead they are applied to temporal intervals, such as disease rates observed for 5- or 10-year intervals of age and period. When the widths of these intervals are equal, the model may be expressed as:

$$E[Y_{ijk}] = \mu + \alpha_i + \pi_j + \gamma_k$$

where μ is the intercept, α_i the effect of age for the i th ($i = 1, \dots, I$) interval, π_j the effect of period for the j th ($j = 1, \dots, J$) interval and γ_k the effect of the k th cohort ($k = i - j + I = 1, \dots, K = I + J - 1$). The usual constraints in this model imply that $\sum \alpha_i = \sum \pi_j = \sum \gamma_k = 0$. The identifiability problem manifests itself through a single unidentifiable parameter (Fienberg and Mason, 1979), which can be more easily seen if we partition each temporal effect into compo-

nents of overall linear trend and curvature or departure from linear trend. For example, age can be given by $\alpha_i = i\beta_a + \tilde{\alpha}_i$, where $i = i - 0.5(I + 1)$, β_a is the overall slope and $\tilde{\alpha}_i$ the curvature. The overall model can be expressed as:



age-period cohort analysis Lexis diagram showing the relationship between age, period and cohort. The diagonal line traces age-period lifetime for an individual born in 1947

$$\begin{aligned} E[Y_{ijk}] &= \mu + (\vec{i}\beta_a + \tilde{\alpha}_i) + (\vec{j}\beta_\pi + \tilde{\pi}_i) \\ &\quad + (\vec{k}\beta_\gamma + \tilde{\gamma}_k) \\ &= \mu + \vec{i}(\beta_a + \beta_\gamma) + \vec{j}(\beta_\pi + \beta_\gamma) + \tilde{\alpha}_i \\ &\quad + \tilde{\pi}_i + \tilde{\gamma}_k \end{aligned}$$

because $\vec{k} = \vec{j} - \vec{i}$. Thus, each of the curvatures can be uniquely determined, but the overall slopes are hopelessly entangled so that only certain combinations can be uniquely estimated (Holford, 1983).

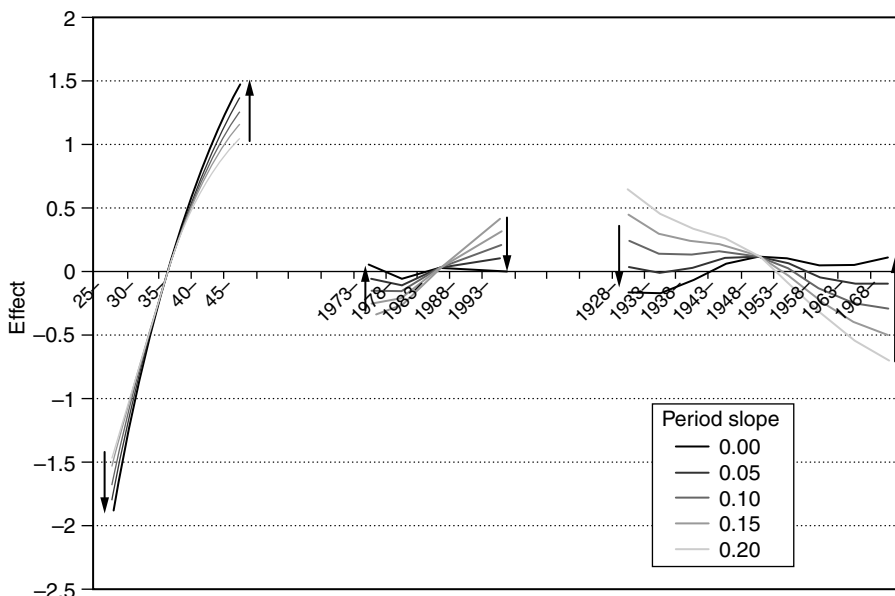
The implication of the identifiability problem is that the overall direction of the effect for any of the three temporal components cannot be determined from a regression analysis. Thus, we cannot even determine whether the trends are increasing or decreasing with cohort, for instance. The second figure on page 8 displays several combinations of age, period and cohort parameters, each set of which provides an identical set of fitted rates. Notice that as the period

parameters are rotated clockwise, the age and cohort parameters are comparably rotated in the counterclockwise direction. Each of these parameters can be rotated a full 180°, but it is important also to realise that they cannot be rotated one at a time, only all together. Thus, even though the specific trends cannot be uniquely estimated, certain combinations of the overall trend can be uniquely determined, such as $\beta_x + \beta_y$, which is called the *net drift* (Clayton and Schifflers, 1987a, 1987b). Alternative drift estimates covering shorter timespans can also be determined and these have practical significance in that they describe the experience of following a particular age group in time, because both period and cohort will advance together. Curvatures, by way of contrast, are completely determined, including polynomial parameters for the square and higher powers, changes in slopes and second differences. The significance test for any one of the temporal effects in the presence of the other two will generally be a test of the corresponding curvature and not the slope. Holford provides further detail on how software can be set up for fitting these models (Holford, 2004). TRH

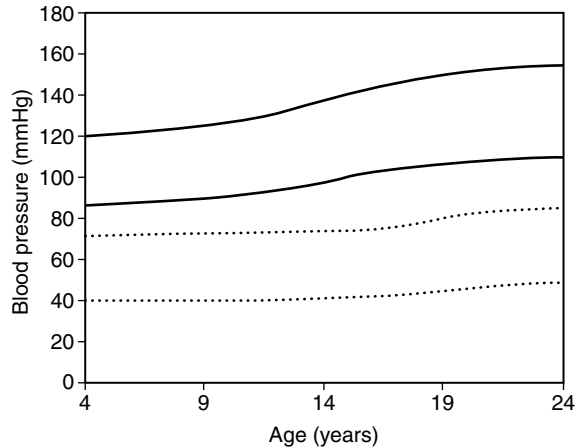
Clayton, D. and Schifflers, E. 1987a: Models for temporal variation in cancer rates I: Age-period and age-cohort models. *Statistics in Medicine* 6, 449–67. **Clayton, D. and Schifflers, E.** 1987b: Models for temporal variation in cancer rates II: Age-period cohort models. *Statistics in Medicine* 6, 469–81. **Fienberg, S. E. and Mason, W. M.** 1979: Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology* 1978, 1-67. **Holford, T. R.** 1983: The estimation of age, period and cohort effects for vital rates. *Biometrics* 39, 311–24. **Holford, T. R.**

2004: Temporal factors in public health surveillance: sorting out age, period and cohort effects. In Brookmeyer, R. and Stroup, D. F. (eds), *Monitoring the health of populations*. Oxford: Oxford University Press, pp. 99–126.

age-related reference ranges These are ranges of values of a measurement that identify the upper and lower limit of normality in the population, where the range varies according to the subject’s age. Reference ranges are an important part of medical diagnosis, where a continuous measurement (e.g. blood pressure) needs converting to a binary variable for decision-making purposes. If the patient’s value lies outside the measurement’s reference range it is treated as abnormal and the patient is investigated further. The construction of reference ranges involves estimating the range of values that covers a specified percentage of the reference population, often 95%. Usually this is the central part of the distribution with equal tail area probabilities, although in some cases the reference range is bounded at zero or infinity. For normally distributed data the range can be derived from the population MEAN and STANDARD DEVIATION (SD), the 95% range, for example, being the mean plus or minus 2 SDs. For nonnormal data the simplest approach is to use quantiles, i.e. rank and count the data, then the 2.5% and 97.5% points are the lower and upper limits of the 95% reference range. However, this is inefficient and requires a large sample. If the data are skew they can be transformed, e.g. to logarithms, and then the reference range can be calculated from the mean and SD on the transformed scale



age-period cohort analysis Age, period and cohort effects for pre-menopausal breast cancer incidence for SEER, 1973–1997



age-related reference ranges Age-related 95 % reference ranges for blood pressure in boys: systolic (solid lines) and diastolic (dotted lines)

and transformed back to the original scale. A more flexible variant is to use a Box–Cox power transformation (of which the logarithm is a special case), which adjusts for skewness more precisely (see TRANSFORMATIONS).

Age-related reference ranges are reference ranges that depend on age. They arise most commonly in paediatrics, notably for age-related measures of body size like height and weight, which can be displayed as GROWTH CHARTS. The principles of reference range estimation are essentially the same when they are age related, except that the ranges for adjacent age groups need to be consistent. To avoid discontinuities at the age group boundaries requires the summary statistics to define the reference range (e.g. the mean and SD) and to change smoothly with age but imposing this constraint complicates the fitting process. For normally distributed homoscedastic data, where the SD is constant across age, the age-related mean can be estimated by LINEAR REGRESSION and the reference range constructed around the regression curve using the residual SD. The regression curve is estimated using a smoothing regression function, e.g. a polynomial, fractional polynomial or generalised additive (cubic spline) curve. If the SD changes with age, as is often the case, a curve of the age-related SD also needs to be estimated by the regression methods of Aitkin (1987) or Altman (1993) and the age-related mean obtained using weighted linear regression with weights corresponding to the inverse square of the age-related SD. The age-related reference range is again constructed around the regression curve using the SD curve.

When the data are skew it may be possible to adjust for the skewness using a single, e.g. logarithmic, transformation at all ages. However, often the degree of skewness is itself age related, although this needs a large sample to show it. In this case an age-related summary statistic for the skewness has to be estimated, along with the age-related mean and SD. The

LMS METHOD is a popular way to do this, or alternatively the EN method of Royston and Wright (1998). For more extreme nonnormal data, a nonparametric approach based on QUANTILE REGRESSION is needed, a form of least absolute errors regression, where smooth curves are constructed for the age-related upper and lower limits of the reference range. The figure gives age-related reference ranges for systolic and diastolic blood pressure in boys aged 4–24, estimated by the LMS method.

There are two advantages of reference ranges based on an underlying frequency distribution, as opposed to those derived using quantile regression. The first is efficiency – the standard errors of the reference range limits are smaller. The second is analytical convenience – data for individuals can be converted to z-SCORES, indicating how many SDs they are above or below the median of the distribution, which is a convenient way of adjusting for age prior to further analysis. *TJC*
[See also GROWTH CHARTS]

Aitkin, M. 1987: Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics* 36, 332–9. **Altman, D. G.** 1993: Construction of age-related reference centiles using absolute residuals. *Statistics in Medicine* 12, 917–24. **Cole, T. J. and Green, P. J.** 1992: Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine* 11, 1305–19. **Koenker, R. W. and D’Orey, V.** 1987: Computing regression quantiles. *Applied Statistics* 36, 383–93. **Royston, P. and Wright, E. M.** 1998: A method for estimating age-specific reference intervals (‘normal ranges’) based on fractional polynomials and exponential transformation. *Journal of the Royal Statistical Society Series A*, 161, 79–101.

age-specific rates These are rates calculated within a number of relatively narrow age bands. A crude rate is the number of events occurring in a population during a specified time period divided by an estimate of the size of the

population. However, when comparing rates between populations with different age distributions, it is necessary to consider rates at specific ages separately.

In the table, death rates are presented for Costa Rica and the United Kingdom for 1999, derived from data from the United Nations (2002). The final column gives the age-specific rates for broad age bands and the crude (total) rate. The age-specific rate is calculated as the number of deaths in the particular age group. In Costa Rica, the death rate at ages 0–5 is calculated as 1296/1 070 000. The rate is expressed per 1000 persons so the rate is multiplied by 1000 to give the rate of 1.2 per 1000 in the final column of the table.

age-specific rates *Population, number of deaths and death rates from all causes for Costa Rica and the United Kingdom for the year 1999*

<i>Costa Rica</i>				
<i>Age group</i>	<i>Population (100 000s)</i>	<i>% in age group</i>	<i>Deaths</i>	<i>Death rate/1000</i>
0–15	10.7	32 %	1296	1.2
15–49	17.4	52 %	2766	1.6
50–69	3.9	12 %	3447	8.8
70+	1.3	4 %	7523	56.6
Total	33.4		15032	4.5
<i>United Kingdom</i>				
<i>Age group</i>	<i>Population (100 000s)</i>	<i>% in age group</i>	<i>Deaths</i>	<i>Death rate/1000</i>
0–15	113.9	19 %	5850	0.5
15–49	288.0	48 %	31 228	1.1
50–69	126.1	21 %	120 759	9.6
70+	67.0	11 %	474 225	70.8
Total	595.0		632 062	10.6

The crude (total) rate for Costa Rica is less than half that for the UK. However, at no age is the rate in the UK double that for Costa Rica and for some age groups the rate is higher in Costa Rica than in the UK. Note that the percentages of the population in each age group (third column) differ markedly. The UK population is much older (11 % of the population are over 70 compared with 4 % in Costa Rica). The different age structure explains the misleading comparison between the crude rates.

Age-specific rates are cumbersome to compare across a number of populations. Standardisation methods are often used to provide an age-adjusted summary rate for each population.

Many countries publish age-specific rates for all cause and specific causes of death, e.g. the annual publications of the

Office of National Statistics (ONS) in England and Wales (ONS, 2002). Age-specific disease incidence rates are also published in various countries, most notably cancer incidence, for which international data are compiled by the International Agency for Research on Cancer (Parkin *et al.*, 2003). Age-specific prevalence rates for exposures such as smoking can also be derived, but are more usually obtained from specific surveys such as the General Household Survey (Walker *et al.*, 2001).

HI

[See also CAUSE-SPECIFIC DEATH RATE, STANDARDISED MORTALITY RATIO]

Office for National Statistics 2002: *Mortality statistics: cause. Review of the Registrar General on deaths by cause, sex and age, in England and Wales, 2001*. London: Office for National Statistics. **Parkin, M., Whelan, S., Ferlay, J., Teppo, L. and Thomas, D. B.** 2003: *Cancer incidence in five continents*, Vol. VIII. Lyon: IARC Scientific Publications. **United Nations** 2002: *2000 demographic yearbook*. New York: United Nations. **Walker, A., Maher, J., Coulthard, M., Goddard, E. and Thomas, M.** 2001: *Living in Britain: results from the 2000 General Household Survey*. London: The Stationery Office.

agreement Agreement in repeated assessments is a fundamental criterion for quality of assessments on rating scales. The use of rating scales and other kinds of ordered classifications of complex qualitative variables is interdisciplinary and unlimited. Rating scale assessments produce *ordinal data*, the ordered categories representing only a rank order of the intensity of a particular variable and not a numerical value in a mathematical sense, although the use of numerical labelling could give a false impression of quantitative data. (see RANK INVARIANCE). The main quality concepts of scale assessments are reliability and validity. Reliability (see MEASUREMENT PRECISION AND RELIABILITY) refers to the extent to which repeated measurements of the same object yield the same result, which means agreement in repeated assessments of various designs. In interrater reliability (see MEASUREMENT PRECISION AND RELIABILITY) studies are made of the level of agreement between observers that classify the same object or individual, and intrarater reliability (see INTRAClass CORRELATION COEFFICIENT) studies refer to agreement in test–retest scale assessments by the same rater.

The frequency distribution of pairs of ordinal data is described in a square CONTINGENCY TABLE (see the figure with parts I, II and III on page 11), and in the case of continuous assessments on a visual analogue scale, VAS, by a scatter plot. The percentage agreement (PA) is a basic agreement measure. When the agreement is unsatisfactory small reasons for disagreement can be evaluated by a statistical method that takes account of the rank-invariant properties of ordinal data and that makes it possible to identify and measure systematic disagreement, when present, separately from disagreement

caused by individual variability in assessments. Systematic disagreement is population based and reveals a systematic change in conditions or memory bias between test–retest assessments, or between raters who interpret the scale categories differently. Large individual variability, on the other hand, is a sign of poor quality of a rating scale as it allows for uncertainty in repeated assessments. The presence of systematic disagreement in the use of the scale categories between the two assessments is revealed by different frequency distributions, which means marginal distributions (parts I and II). A systematic disagreement regarding the categorical levels and in the way of concentrating the assessments on the categories are measured by the relative position (RP) and the relative concentration (RC) respectively. The RP expresses the extent to which the marginal distribution of assessments Y is shifted towards higher categories than the marginal distribution of X , rather than the opposite. A theoretical description is the difference between the probabilities $P(X < Y) - P(Y < X)$. Possible values of RP range from (-1) to 1 , and RP is positive when higher scale categories are more frequently used in the assessments Y than in X when compared with the opposite. Correspondingly, the RC expresses the extent to which the marginal distribution of Y assessments is more concentrated to central scale categories than is the marginal distribution of X , theoretically described by the difference in probabilities $P(X_i < Y_k < X_j) - P(Y_i < X_k < Y_j)$. Possible values range from (-1) to 1 , and a positive RC indicates that the assessments Y are more concentrated than X . Zero or very small values of both RP and RC mean that the systematic part of an observed disagreement paired assessments is negligible.

Systematic disagreement is evident by the marginal heterogeneity, and by pairing off the two sets of marginal frequencies, the so-called rank-transformable pattern of agreement (RTPA) is constructed. The RTPA describes the expected pattern in the case of systematic disagreement only. All pairs of observations of the RTPA will have the same rank ordering in the two assessments provided that the ranks are tied to the cells, which is the definition of the augmented ranking procedure (aug-ranks) (see RANKING).

Part II in the figure is the RTPA of the pattern in part I. The observed distribution of pairs in part I deviates from this RTPA, which means that some of the pairs of aug-ranks given to the observations differ. The relative rank variance (RV) is a rank-based measure of this observed individual variability, i.e. unexplained by the measures of systematic disagreement:

$$RV = \frac{6}{n^3} \sum_{i=1}^m \sum_{j=1}^m x_{ij} (\Delta \bar{R}_{ij})^2$$

where n is the number of paired assessments and $(\Delta \bar{R}_{ij})^2$ is the square of the mean aug-rank difference of the ij th cell, and the summation is made over all cells ij of the $m \times m$ square table, $0 \leq RV \leq 1$ (Svensson *et al.*, 1996; Svensson, 1998a). The Cohen’s coefficient kappa (κ) is a commonly used measure of agreement adjusted for the chance expected agreement (see KAPPA AND WEIGHTED KAPPA).

The calculations of Cronbach’s alfa and other so-called reliability coefficients are based on the assumption of quantitative, normally distributed data, which is not achievable in

I	Rater X					II	Rater X					III	Rater X					
	A	B	C	D	tot		A	B	C	D	tot		A	B	C	D	tot	
Rater Y	D			1	1	2	D				2	2	D		1	4	14	19
	C		2	2	14	18	C			1	17	18	C	1	2	10	4	17
	B	1	1	11	3	16	B			16		16	B	1	6	3	1	11
	A	2	8	3	1	14	A	3	11			14	A	1	2			3
tot	3	11	17	19	50	tot	3	11	17	19	50	tot	3	11	17	19	50	

PA, 12 %
 RP, -0.49 RC, 0.16
 RV, 0.08

PA, 12 %
 RP, -0.49 RC, 0.16
 RV, 0

PA, 62 %
 RP = RC = 0
 RV, 0.05

agreement Examples of paired ordinal data from interrater assessments on a four-point scale with the ordered categories labelled $A < B < C < D$. The rank-transformable pattern of agreement (RTPA) is shaded. The measures of percentage agreement (PA), the relative position (RP), the relative concentration (RC) and the relative rank variance (RV) are given

data from rating scales. There is also a widespread misuse of the correlation coefficient as a reliability measure. The correlation coefficient (see CORRELATION) measures the degree of association between two variables and does not measure the level of agreement. In part I of the figure the PA is 12 %, and the observed disagreement is mainly explained by a systematic disagreement in position. The negative RP value (−0.49) and the RTPA (part II) shows that the assessments Y systematically used lower categories than X . A slight additional individual variability, $RV = 0.08$ is observed. SPEARMAN'S RANK CORRELATION COEFFICIENT, r_s , is 0.66 in part I of the figure and 0.97 in part II, ignoring the fact that the assessments are systematically biased and unreliable. The same holds for the coefficient kappa (−0.14). In part III the marginal homogeneity and the zero RP and RC values confirm that the disagreement (39 %) is entirely explained by slight individual dispersion ($RV = 0.05$) from the RTPA, which is the main diagonal in this case. The r_s is 0.61 and the κ is 0.45.

Besides reliability studies, the level of disagreement is of main interest in paired assessments 'before and after' treatment for analysing change in outcome or treatment effect. In this application of the disagreement measures, nonzero RP and RC values indicate the level of common group change in outcomes, and the heterogeneity in changes among the individuals is measured by the RV (Svensson, 1998b). ES

Svensson, E. 1998a: Application of a rank-invariant method to evaluate reliability of ordered categorical assessments. *Journal of Epidemiology and Biostatistics* 3, 403–9. **Svensson, E.** 1998b: Ordinal invariant measures for individual and group changes in ordered categorical data. *Statistics in Medicine* 17, 2923–36. **Svensson, E., Starmark, J.-E., Ekholm, S., von Essen, C. and Johansson, A.** 1996: Analysis of inter-observer disagreement in the assessment of subarachnoid blood and acute hydrocephalus on CT scans. *Neurological Research* 18, 487–94.

Akaike's information criterion Akaike's information criterion (AIC) is an index used to discriminate between competing models. It is widely used when there is the issue of model choice where we wish to find the most parsimonious model (see Akaike, 1974). Often there may be a number of possible models that can be fitted to the data, from which parameters can be estimated using, for example, the MAXIMUM LIKELIHOOD ESTIMATION. Generally, complex models are more flexible, but contain a relatively large number of parameters, whereas simpler models with fewer parameters may compromise the fit of the model to the data. Essentially, the AIC statistic compares competing models by considering the trade-off between the complexity of the model and the corresponding fit of the model to the data. The AIC statistic is widely used, particularly as it can be used to compare even

nonnested models when likelihood ratio tests cannot be applied.

Let \mathbf{x} denote the data and $\hat{\theta}$ the corresponding maximum likelihood estimates (MLEs) of the parameters. Then, the AIC for a given model is denoted by:

$$AIC = -2 \log L(\hat{\theta}; \mathbf{x}) + 2p$$

where p denotes the number of parameters in the given model being fitted to the data and $\log L(\hat{\theta}; \mathbf{x})$ the corresponding log-likelihood evaluated at the MLEs of the parameters. The AIC statistic is calculated for each possible model being considered. The model deemed optimal is the one with the smallest AIC value, i.e. a model with a relatively small number of parameters that adequately fits the data. The AIC is generally easy to calculate given the maximum of the likelihood function and is very versatile, allowing us to compare, for example, nonnested models. We note that corrections have been suggested to the AIC statistic to allow for data with overdispersion (denoted by QAIC) and small sample sizes (AIC_c). See, for example, Burnham and Anderson (2002), Sections 2.4–5.

The AIC statistic has also been used to compare the performance of different models, relative to each other (Buckland, Burnham and Augustin, 1997; Burnham and Anderson, 2002, Section 2.6). It is not the absolute values of the AIC statistics that are important but their relative values, in particular their difference. For each model the term $\Delta AIC = AIC - \min AIC$ is calculated, where $\min AIC$ is the value of the AIC statistic for the model deemed optimal. Clearly, $\Delta AIC = 0$ for the model deemed optimal; the larger the value of ΔAIC the poorer the model. The relative penalised likelihood weights w_i can also be calculated for each model $i = 1, \dots, m$, where:

$$w_i = \frac{\exp(-\Delta AIC_i/2)}{\sum_{j=1}^m \exp(-\Delta AIC_j/2)}$$

and AIC_i denotes the corresponding AIC value associated with model i . The weights provide a scale to interpret the difference in values for the models. Finally, these model weights can be used to obtain a (weighted) model-averaged estimate of parameters of interest. RK

[See also DEVIANCE, LIKELIHOOD RATIO]

Akaike, H. 1974: A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC 19, 716–72. **Buckland, S. T., Burnham, K. P. and Augustin, N. H.** 1997: Model selection: an integral part of inference. *Biometrics* 53, 603–18. **Burnham, K. P. and Anderson, D. R.** 2002: *Model selection and multimodel inference*, 2nd edition. Heidelberg: Springer Verlag.

allelic association This is an association between two alleles (at two different loci), or between an allele and a phenotypic trait, in the population. Since humans are diploid a more technical definition of the former is necessary: two alleles are associated if their frequency of co-occurrence in the same haplotype (i.e. the genetic material transmitted from one parent) is greater than the product of the marginal frequencies of the two alleles.

Association between two alleles is also known as *linkage disequilibrium*. The reason is that, in a large population under random mating, the extent of association between two alleles (as measured by the difference between the frequency of the haplotype containing the two alleles and the product of the frequencies of the two alleles) decreases by a factor equal to one minus the recombination fraction (see GENETIC LINKAGE) between the two loci, per generation. Thus allelic association represents a state of disequilibrium that tends to dissipate at a rate determined by the strength of linkage between the two alleles, towards the state of equilibrium when the frequency of the haplotype is equal to the product of the frequencies of the two constituent alleles.

Associations between two alleles can arise in a population for a number of reasons. The mutation that gave rise to the more recent allele may have occurred on a chromosome that happened to contain the other allele. Random genetic drift during a population bottleneck may have led to the overrepresentation of some haplotypes. The mixing of two populations with different allele frequencies may have resulted in associations between alleles in the overall population. When, for any of these reasons, such allelic associations arose many generations ago, only those occurring between tightly linked loci are likely to have persisted to the current generation. We would therefore expect an imperfect inverse relationship between the extent of association between two alleles and the distance between them.

An association between an allele and a disease may be the result of a direct causal relationship. In other words, the allele is a causal variant that is functional and increases the risk of the disease. However, it could also be indirect, with the allele being in linkage disequilibrium with a causal variant. The presence of linkage disequilibrium between tightly linked loci means that it is possible to screen a chromosomal region for a causal variant without examining all the alleles, only a sufficient number to ensure that any causal variant in the region is likely to be in linkage disequilibrium with one or more of the alleles examined. The polymorphisms chosen to represent itself and associated polymorphisms in its vicinity in an association study are called TAG polymorphisms. The International HapMap Project (www.hapmap.org) has characterised the pattern of allelic associations among over 3 million single nucleotide polymorphisms (SNPs) in the human genome in three major populations (Europeans, Africans and Asians).

Classical epidemiological designs (CASE-CONTROL STUDIES, COHORT STUDIES, CROSS-SECTIONAL STUDIES) are readily applicable to the study of disease-allele associations, as are the statistical methods developed for these designs (e.g. LOGISTIC REGRESSION, SURVIVAL ANALYSIS). These designs are potentially susceptible to the problem of hidden population stratification, which can lead to spurious associations or mask true associations. Family-based association designs are robust to population stratification and usually consist of the use of either parental or sibling controls. Methods for the analysis of matched samples, such as the McNEMAR'S TEST (also called the transmission disequilibrium test in the context of parental controls) and CONDITIONAL LOGISTIC REGRESSION are applicable to these designs.

The study of disease-allele associations is a complementary strategy to linkage analysis, in the localisation and identification of genes that increase the risk of disease. In general, allelic association is unlikely to be detected when the marker locus is quite far (>1 megabase) from the disease locus, but can be much more powerful than linkage when the marker locus is close enough to the disease locus to be in substantial linkage disequilibrium with it, particularly when the effect size of the disease locus is small. For this reason, allelic association is particularly appealing for searching regions that demonstrate linkage to the disease or to the investigation of specific candidate genes. However, technological developments have enabled the efficient genotyping of up to 1 million SNPs in a single array, and this has led to association studies on the whole-genome scale (called genome-wide association studies, or GWAS) that have coverage of over 90% of common variants (allele frequency $> 5\%$) in the genome. PS

all subsets regression A form of regression in which *all* possible models are compared using some appropriate criterion for indicating the 'best' models. If there are p explanatory variables in the data, there are a total of $2^p - 1$ possible regression models because each explanatory variable can be in or out of the model and the model containing no explanatory variables is excluded. One possible criterion for comparing models is the MALLOW'S C_p STATISTIC and to illustrate its use we will apply it to data that arise from a study of 25 patients with cystic fibrosis reported in O'Neill *et al.* (1983), and also given in Altman (1991). Data for the first three patients are given in the first table. The dependent variable in this case is a measure of malnutrition (PE_{\max}). Some of the models considered in the all subsets regression of these data are shown in the second table, together with their associated C_p values, where p refers to the number of parameters in a particular model, i.e. a model that includes a subset of $p - 1$ of the explanatory variables plus an intercept. If C_p is plotted against p , the subsets of explanatory

all subsets regression *Cystic fibrosis data; first three subjects*

Sub	Age	Sex	Height	Weight	BMP	FEV	RV	FRC	TLC	PE _{max}
1	7	0	109	13.1	68	32	258	183	137	95
2	7	1	112	12.9	65	19	449	245	134	85
3	8	0	124	14.1	65	22	441	268	147	100

Sub: subject number
 Sex: 0 = male, 1 = female
 BMP: body mass (weight/height²) as a percentage of the age-specific median in normal individuals
 FEV: forced expiratory volume in one second
 RV: residual volume
 FRC: functional residual capacity
 TLC: total lung capacity
 PE_{max}: maximal static expiratory pressure (cmH₂O)

all subsets regression *Some of the models fitted in applying the all subsets regression to the cystic fibrosis data (size is one more than the number of variables in a model, to include the intercept)*

Model	Size	Terms	C _p
7	2	Sex	17.24
14	3	Sex, weight	4.63
21	4	Age, FEV, RV	2.62
28*	4	Age, BMP, FEV	4.5
35	5	Sex, weight, BMP, FEV	2.95
42	6	Age, weight, BMP, FEV, RV	2.8
49	6	Age, sex, height, FEV, TLC	6.99
56*	7	Age, sex, height, FEV, RV, TLC	7.06
63	8	Sex, weight, BMP, FEV, RV, FRC, TLC	6.49
70	9	Age, height, weight, BMP, FEV, RV, FRC, TLC	8.06
77	9	Age, sex, height, BMP, FEV, RV, FRC, TLC	10.29

* Models close to the line C_p = p.

variables most worth considering in trying to find a parsimonious model are those lying close to the line C_p = p.

All subsets regression has been found to be particularly useful in applications of COX'S REGRESSION MODEL (see Kuk, 1984). BSE
 [See also MULTIPLE LINEAR REGRESSION]

Altman, D. G. 1991: *Practical statistics for medical research*. London: CRC/Chapman & Hall. Kuk, A. Y. C. 1984: All subsets regression in a proportional hazards model, *Biometrika*, 71, 587-92. O'Neill, S., Leahy, F., Pasterkamp, H. and Tal, A. 1983: The effects of chronic hyperfunction, nutritional status and posture on respiratory muscle strength in cystic fibrosis. *American Review of Respiratory Disorders* 128, 1051-4.

alternative hypothesis See HYPOTHESIS TESTS

AMOS See STRUCTURAL EQUATION MODELLING SOFTWARE

analysis of covariance (ANCOVA, ANOCOVA)

This is an extension of the analysis of variance (ANOVA) that incorporates a continuous explanatory variable. Where ANOVA aims to detect if there is a change in the mean value of a variable across two or more groups, ANCOVA (or rarely ANOCOVA) does the same but adjusts for a continuous covariate.

Most commonly this covariate will be a baseline measurement, allowing the analysis to adjust for initial variation between participants and isolate the effects due to the treatment factor. However, sometimes a different covariate is used. For example, Karhune *et al.* (1994) consider the association between alcohol intake (divided into four categories) and numbers of Purkinje cells. In doing so they introduce age as a continuous covariate in order to 'control' or 'adjust' for the effects of age on cell numbers.

Under other circumstances the authors could have been interested in the effects of age and wanting to adjust for alcohol intake. Despite being the same analysis computationally, this is not typically what is thought of as analysis of covariance and might more commonly be presented as a 'regression'. Indeed the various analysis of variance methods can all be viewed from within a regression framework, which demonstrates that ANCOVA can be extended to cope with much more than one continuous covariate.

Mathematically, ANCOVA follows a similar path to that for ANOVA and the output is usually summarised in a similar table, although the details may vary.

The promised benefits of the analysis of covariance are clear. If one has an unbalanced observational study, then ANCOVA can adjust for differences in baseline values and remove a potential bias from the results. By the same token, if one has a randomised trial that is naturally balanced, then

ANCOVA reduces the amount of unexplained variation in the data and thus increases the power of the test.

However, ANCOVA can only be employed if the appropriate assumptions are met. These include those of ANOVA (i.e. normality of residuals, homoscedasticity) as well as the appropriateness of the ANCOVA model. Is the relationship with the covariate truly linear? Does the effect of the covariate vary between groups? Failing to meet these assumptions can lead to the introduction of important but subtle biases. It is a frequent concern that medical research papers report a covariate as having been ‘controlled’ or ‘adjusted’ for, with no evidence that the control or adjustment was appropriate. For further details see Altman (1991), Owen and Froman (1998), Miller and Chapman (2001) and Vickers and Altman (2001).

AGL

[See also GENERALISED LINEAR MODEL]

Altman, D.G. 1991: *Practical statistics for medical research*. London: Chapman & Hall. **Karhune, P. J., Erkinjuttu, T. and Laippala, P.** 1994: Moderate alcohol consumption and loss of cerebellar Purkinje cells. *British Medical Journal* 308, 1663–7. **Miller, G. A. and Chapman, J. P.** 2001: Misunderstanding analysis of covariance. *Journal of Abnormal Psychology* 110, 40–8. **Owen, S. V. and Froman, R. D.** 1998: Uses and abuses of the analysis of covariance. *Research in Nursing and Health* 21, 557–62. **Vickers, A. J. and Altman, D. G.** 2001: Analysing controlled trials with baseline and follow-up measurements. *British Medical Journal* 323, 123–4.

analysis of variance (ANOVA) Often referring to the one-way analysis of variance, it is a test for a common MEAN in multiple groups that we describe in detail here. Analysis of variance frequently arises in the comparison of more complicated models, but the same logical arguments apply. In all cases, the underlying concept is to partition the observed variance into quantities attributable to specific explanatory sources, and then consider important those sources that explain ‘more than their fair share’ of the variance.

Despite the confusion sometimes caused by the name, the one-way analysis of variance is a method for testing to see whether multiple samples come from populations that share the same mean. In this respect it can be viewed as an extension to the *t*-test, which assesses whether samples from two populations share a common mean. An analysis of variance performed on two samples is equivalent to performing a *t*-test.

ANOVA assumes that all the samples come from populations with a NORMAL DISTRIBUTION that share the same VARIANCE. It can be viewed in a number of ways, but essentially compares the estimate of the variance obtained within samples (that makes no assumption that the populations have a common mean) with an estimate of the variance from the sample means (which will require the assumption that

the populations have the same mean). If the two estimates of the variance are different, then this is evidence that our assumption of equality failed and, therefore, that the populations do not all have the same mean.

Note that the variance of a single sample is estimated as the sum of squared differences from the mean divided by the sample size minus one. The sum of squared differences term is interpretable as a measure of the total variation in the sample. In the analysis of variance, by combining all groups together, one can calculate this measure for all the data. This is termed the ‘total sum of squares’ or ‘total SS’.

Variation in the data is either ‘between’ or ‘within’ the samples. The ‘within groups sum of squares’ or ‘within SS’ can be calculated as the sum of squared differences from the individual sample means (rather than the differences from the overall mean that produced the total SS). ‘Between groups sum of squares’ or ‘between SS’ can be calculated directly, but is most easily calculated by subtraction of the within SS from the total SS.

The two estimates of the variance (or ‘mean square’ as it is often termed in this context) can then be calculated. The between groups mean square is equal to the between SS divided by the number of groups minus one. The within groups mean square is equal to the within SS divided by the number of observations minus the number of groups.

An *F*-statistic is then calculated as the between groups variance divided by the within groups variance. Under the assumptions of normality and homoscedasticity (common variance) this statistic will be an observation from an *F*-DISTRIBUTION if the groups come from populations with a common mean. The DEGREES OF FREEDOM of the *F*-distribution are the number of groups minus one and the number of observations minus the number of groups.

From the *F*-distribution, we can calculate the probability of observing such an extreme value of the *F*-statistic if the populations have a common mean. This is a one-tailed test. If the value is unusually small, this suggests the between groups variance is unusually small and so is not evidence of variation between the groups. Therefore, the test is to find the probability, if the populations do have a common mean, of observing a value greater than that observed.

A natural way of presenting ANOVA is the ANOVA table. Given *N* observations that fall into *k* groups, it is necessary to calculate the total SS and the within SS as described earlier and then the analysis can be completed as presented in the first table. Murphy *et al.* (1994) conducted an analysis of variance to see if milk consumption before the age of 25 affects bone density of the hip in later life. A total of 248 women participated in this part of their study (*N* = 248) and were divided into groups that represent low, medium and high milk consumptions (*k* = 3). The samples had similar variances and so at least one of the assumptions for ANOVA was

analysis of variance (ANOVA) *The analysis of variance table*

Source of variance	Degrees of freedom	Sums of squares	Mean squares	F	P
Between groups	$k - 1$	Between SS = Total SS – Within SS	Between MS = Between SS/($k - 1$)	$\frac{\text{Between MS}}{\text{Within MS}}$	p
Within groups	$N - k$	Within SS	Within MS = Within SS/($N - k$)		
Total	$N - 1$	Total SS			

analysis of variance (ANOVA) *Approximate reconstruction of the analysis of variance table from Murphy et al. (1994).*

Source of variance	Degrees of freedom	Sums of squares	Mean squares	F	P
Between groups	2	0.15	0.08	3.8	0.23
Within groups	245	4.4	0.02		
Total	247	4.6			

(Entries in bold were inferred from the paper, the rest simply follow from the calculations)

satisfied. As is common for reasons of space, the ANOVA table was not presented in the published paper, just the P -value, but enough data were presented for an approximate reconstruction.

We can infer that the within SS is approximately 4.4 and the between SS is approximately 0.15. This leads to an F -statistic of approximately 4. From the reported P -value (0.023), it can be calculated from the F -distribution (with 2 and 245 respectively for numerator and denominator degrees of freedom) that the F -statistic was 3.8. The conclusion then is that there is evidence that these samples do not come from populations that share a common mean. The reconstructed table is presented in the second table (entries in bold in this table were inferred from the paper, the rest simply follow from the calculations).

It is preferable to conduct an analysis of variance rather than to conduct t -tests between all pairs of groups. ANOVA avoids problems of multiple testing and thus keeps control of the SIGNIFICANCE LEVEL. Having conducted an ANOVA and rejected the hypothesis of common means, it may then be desired to test to see which groups are responsible (although a plot of the data might be as informative). In this case, care must be taken to correct for the problems of making MULTIPLE COMPARISONS.

It is important to take note of the assumptions being made, rather than simply ignoring them. ANOVA can be quite robust to variations from normality, but heteroscedasticity can be a serious problem. Residual plots can be used to help assess the normality and BOXPLOTS can be used to help assess the heteroscedasticity. Possible formal tests for the assumptions are the KOLMOGOROV-SMIRNOV TEST and LEVENES TEST respectively.

If the assumptions do not hold, then TRANSFORMATION of the data might correct this. Otherwise a number of nonparametric alternatives to ANOVA exist, the most commonly used being the KRUSKAL-WALLIS TEST and the FRIEDMAN TEST.

The one-way analysis of variance is appropriate when our data are simply divided into a number of groups. There are many other forms of analysis of variance. The TWO-WAY ANALYSIS OF VARIANCE should be used when the groups are defined by two factors. Suppose, for example, we had six groups: the three groups of women in Murphy et al. (1994) and three groups of men at the same levels of milk consumption. Rather than a one-way analysis of variance, a two-way analysis of variance with gender and milk consumption as the two factors would be appropriate in this instance.

If the data are multiple observations from the same subjects, perhaps measurements of cholesterol levels 0, 7, 14, 21 and 28 days after starting a new diet on several individuals, then a REPEATED MEASURES ANALYSIS OF VARIANCE would be appropriate. This is a special case of the two-way ANOVA and can be viewed as an extension of the paired sample t -test.

If there are observations of more than one characteristic from the individuals in several groups, i.e. measures of both the diastolic and systolic blood pressure, then a multivariate analysis of variance (MANOVA) can be used. If, however, it is desired to correct for a measured baseline covariate, such as body mass index, in the analysis, then an ANALYSIS OF COVARIANCE (ANCOVA) may be used.

All these techniques could be implemented through a regression framework, in most cases MULTIPLE LINEAR REGRESSION. The advantages of doing so would be the transition from the use of a HYPOTHESIS TEST to an actual estimate of effect

sizes. This approach would also allow more flexibility; for instance in the case of Murphy *et al.* (1994) we could account for the natural ordering of the levels of milk consumption that ANOVA ignores. As a general principle, estimation and modelling are usually preferred to testing of hypotheses. For further details see Altman (1991) and Altman and Bland (1996). AGL

Altman, D. G. 1991: *Practical statistics for medical research*. London: Chapman & Hall. **Altman, D. G. and Bland, J. M.** 1996: Statistics notes: comparing several groups using analysis of variance. *British Medical Journal* 312, 1472–3. **Murphy, S., Khaw, K.-T., May, H. and Compston, J. E.** 1994: Milk consumption and bone mineral density in middle aged and elderly women. *British Medical Journal* 308, 939–41.

area under the curve (AUC) This is a simple and useful method of obtaining a summary measure from plotted data. Medical research is frequently concerned with serial data, as in repeated measurements (see REPEATED MEASURES ANALYSIS OF VARIANCE) on a subject over time, e.g. blood aspirin concentration measured at various times over a 2-hour interval (Matthews *et al.*, 1990). Say we have n measurements y_i taken at times t_i ($i = 1, \dots, n$). Such data are frequently exhibited by plotting y_i versus t_i and joining the resulting points by straight-line segments resulting in a ‘curve’. The resulting area under the curve (AUC) is often used as a single-number summary measure for the individual subject. Further analysis of the subjects or comparison of groups of subjects is carried out based on the summary measures. The AUC for the set of points (y_i, t_i) $i = 1, \dots, n$ is typically calculated by the trapezium rule:

$$\text{AUC} = \frac{1}{2} \sum_{i=1}^{n-1} (t_{i+1} - t_i)(y_i + y_{i+1})$$

The AUC is used as a summary measure in many areas of medical research, including bioequivalence and pharmacokinetics. It plays an especially important role in the analysis of RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES. The area under the ROC curve of a diagnostic marker (test) measures the ability of the marker to discriminate between healthy and diseased subjects. It is the most commonly used measure of performance of a marker. We use the convention that larger marker values are more indicative of disease. Then if we randomly pick one subject from the healthy population and one from the diseased population we would ‘expect’ that the value of the marker for the healthy subject would be smaller than the corresponding value for the diseased subject. AUC is the probability that this, in fact, occurs. The larger the AUC, the better the overall discriminatory accuracy of the marker. An area of 1 represents a perfect test while an area of 1/2 represents a worthless test having a discriminatory ability, which is the equivalent of differentiating between healthy and

diseased subjects by a fair coin toss. Consider the example discussed in the entry for the ROC curve. The points on the curve are given in the table.

area under the curve *Summary data used in an ROC curve*

Specificity	(y_i)	0	0.56	0.84	0.94	0.98	1.00
1-Sensitivity	(t_i)	0	0.04	0.12	0.32	0.60	1.00

The data presented result in an AUC as follows:

$$\begin{aligned} \text{AUC} &= 0.5[(0.04-0) \times (0 + 0.56) + (0.12-0.04) \\ &\quad \times (0.56 + 0.84) + (0.32-0.12) \times (0.84 \\ &\quad + 0.94) + (0.60-0.32) \times (0.94 + 0.98) \\ &\quad + (1.00-0.60) \times (0.98 + 1.00)] = 0.91 \end{aligned}$$

An area of 0.91 indicates the high discriminatory ability of the marker.

For the ROC curve, estimating the area by the trapezium rule is equivalent to computing the Wilcoxon or Mann–Whitney statistic divided by the products of the sample sizes on the healthy and diseased populations. For smoothed ROC curves, alternative estimates of the AUC are available (Faraggi and Reiser, 2002). The effectiveness of alternative diagnostic markers is usually studied by comparing their AUCs (Wieand *et al.*, 1989). Adjustments of these areas for covariate information, selection bias and pooling effects are discussed in the references given in the entry for the ROC curve. Schisterman *et al.* (2001) consider corrections of the AUC for measurement error. For further details see Hanley and McNeil (1982). DF/BR

Faraggi, D. and Reiser, B. 2002: Estimation of the area under the ROC curve. *Statistics in Medicine* 21, 3093–106. **Hanley, J. A. and McNeil, B. J.** 1982: The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36. **Matthews, J. N. S., Altman D. G., Campbell, M. J. and Royston, P.** 1990: Analysis of serial measurements in medical research. *British Medical Journal* 300, 230–5. **Schisterman, E., Faraggi, D., Reiser, B. and Trevisan, M.** 2001: Statistical inference for the area under the ROC curve in the presence of random measurement error. *American Journal of Epidemiology* 154, 174–9. **Wieand, S., Gail, M. H., James, B. R. and James, K. L.** 1989: A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 76, 585–92.

artificial intelligence (AI) This branch of computer science is devoted to the simulation of intelligent behaviour in machines. Traditional focus areas of AI are machine vision, MACHINE LEARNING, natural-language processing and speech recognition. Historically an interdisciplinary field, and hence characterised by the presence of several

competing paradigms and approaches, recently AI has started developing a more unified conceptual framework, based largely on the convergence of statistical and algorithmic ideas.

A constant theme of AI throughout its history has been 'pattern recognition', the crucial task of detecting 'patterns' (regularities, relations, laws) within data. This task has emerged as a roadblock in all the traditional areas mentioned earlier and hence has attracted significant attention. Since most current approaches to pattern recognition involve significant use of statistics, this has become an important tool in AI in general.

Recently, AI has been applied to a new series of important problems and this, in turn, has heavily affected general AI research. Important applications of modern AI include: intelligent data analysis (see also DATA MINING IN MEDICINE); information retrieval and filtering from the web; bioinformatics; and computational biology. Traditional application areas, by way of contrast, included the design of EXPERT SYSTEMS for medical or industrial diagnosis, methods for scheduling in logistics and creation of other decision-making assistant software.

The imprecise definition of what AI actually is has made it harder in time to gauge the impact of this research field on everyday applications. A number of widely used computer programs would have met early definitions of artificial intelligence, e.g. popular web-based recommendation systems or air travel planning advisors.

Popular techniques for pattern recognition such as NEURAL NETWORKS, decision trees and cluster analysis (see CLUSTER ANALYSIS IN MEDICINE) have made their way into the standard toolbox of data analysis and are commonly found in the toolbox of any biology lab. Machine vision methods are routinely used in analysing medical images, as well as parts of systems such as microarray machines for collecting gene expression data. Web retrieval and email filtering software also incorporate several ideas from natural-language processing and pattern recognition and the modern sequence analysis of genomic data heavily relies on techniques originally developed for speech recognition. Intelligent web agents exist to find, assess and retrieve relevant information for the user and speech-recognition systems are routinely used in automatic phone information systems. The field of artificial intelligence has clearly produced a number of practical applications, but – the critics say – these have been achieved without solving the general problem of building intelligent machines. Maybe for this reason, generally the main success story of AI is reported to be the defeat of the chess world champion Gary Kasparov by an IBM algorithm in 1997.

The origin of the field of AI is often identified with a paper by A. M. Turing, which appeared in 1950 in the journal *Mind*, and with a workshop held at Dartmouth College in the

summer of 1956, although many key ideas had already been debated before, during the early years of cybernetics.

Modern techniques of artificial intelligence include Bayesian belief networks, part of the more general field of probabilistic graphical models; pattern-recognition algorithms such as SUPPORT VECTOR MACHINES, which represent the convergence of ideas from classical statistics and from neural networks analysis; statistical analysis of natural language text and machine vision algorithms; reinforcement learning algorithms which represent a connection with control theory; and many other methods. *NC/TDB*

Bishop, C. 1996: *Neural networks for pattern recognition*. Oxford: Oxford University Press. **Mitchell, T.** 1995: *Machine learning*. Maidenhead: McGraw-Hill. **Russell, S. and Norvig, P.** 2002: *Artificial intelligence: a modern approach*, 2nd edition. Harlow: Prentice Hall. **Shawe-Taylor, J. and Cristianini, N.** 2004: *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.

association This is the statistical dependence between two variables. Measures of association, unlike descriptive statistics of a single variable, summarise the extent to which one variable increases or decreases in relation to a change in a second variable. The basic graphical analysis of two variables is the SCATTERPLOT, which provides evidence of association in the shape and direction of the scatter of points. In the example given here, there appears to be an association between body mass index and systolic blood pressure values in a sample of a few thousand middle-aged men and women: higher values of body mass index tend to be associated with higher values of systolic blood pressure, suggesting a 'positive' association. A 'negative' association, in contrast, would describe a situation where an increase in one variable tends to be related to a decrease in the second variable.

Various statistical measures can be used to interpret the degree of association.

Correlation coefficient. This specifically measures the degree of *linear* association between two quantitative variables on a scale from negative one to positive one. A value of zero indicates a total absence of linear association, while a value of positive or negative one indicates a perfect linear relationship. The correlation coefficient between body mass index and systolic blood pressure in our example was 0.25, indicating a positive association that is less than perfectly linear. However, adherence to a linear relationship is only one form of association and it is easy to imagine other plausible patterns of association, such as a parabolic scatter, in which the change in one variable may be perfectly reflected in the change in the second variable, but the correlation coefficient might be close to zero.

Regression coefficient. In the case of simple linear regression, there is a complete correspondence between the correlation coefficient and the regression coefficient for the slope (β). The regression coefficient, therefore, also measures association, but its value is interpreted as the magnitude of change in the dependent variable that arises, on average, from a unit change in the independent variable. In our example, an estimate of $\hat{\beta} = 1.37$ indicated that a 1 kg/m^2 increase in the body mass index was associated with an average increase of 1.37 mmHg systolic blood pressure. However, in more complex regression models, the regression coefficient can measure other forms of association beyond linear dependence. For example, either the dependent or independent variable may be mathematically transformed, such as raising to a higher power, taking logarithms, etc., and the association measured by the regression coefficient would express a nonlinear change in one variable in response to a change in the second variable.

Relative risk. In the special case of two binary variables, various ratio measures are often used to quantify the degree of association. For example, one variable might be a measure of disease occurrence, the other a biological or environmental quantity. Most commonly the ratio would compare probability of disease expressed as an odds, a risk or some other relevant approximation to the risk. A relative risk value of 1, indicating equal risks in both groups, suggests that no association exists between the biological or environmental quantity and disease.

If a statistical measure suggests positive or negative association, this should not immediately be taken to imply that the association is valid and generalisable. Several considerations might lead us to question the importance of an observed statistical association.

First, consideration of the STANDARD ERROR of the measure of association, generally reflecting the size of the sample, places the magnitude of association in perspective with the magnitude of random error. Apparently strong associations may in fact be poorly estimated and fall short of statistical significance.

Second, an apparent association may be entirely spurious (i.e. 'confounded') due to the influence of other measured, or unmeasured, variables that have not been accounted for in the analysis. For example, in a preliminary statistical enquiry, risk of coronary heart disease may appear to be associated with watching television, although consideration of the underlying relationship with obesity and physical exercise would probably suggest that the preliminary finding was spurious. An association may alter after adjustment for the interdependence of other variables and the general validity of a measure of association would often depend on the extent to which such potential interdependencies have been taken into account. Studies measuring several variables often utilise

multiple regression models to estimate adjusted regression coefficients and partial correlation coefficients by including all relevant variables in the model. However, even after allowing for such interdependencies, the much stronger claim of CAUSALITY between two variables would generally require examination of more stringent criteria.

Third, an observed association may be specific to the chosen range of the variables or to the particular group of subjects studied and any inference beyond the range of the data to hand would require careful consideration of the method of sample selection. Various forms of selection bias may limit the generalisability of the association. JGW

as treated See INTENTION-TO-TREAT

attenuation due to measurement error This is a bias reducing the size of a correlation or a regression coefficient due to imprecision of data measurement. Consider an analytical epidemiological study in which the aim is to estimate the CORRELATION between true average consumption of alcohol (mg per day) and true average systolic blood pressure (mmHg). Blood pressure measurements are well-known to be variable within individuals and a single measurement is likely to be rather imprecise (see MEASUREMENT PRECISION AND RELIABILITY). Such a statement is even more true of a single day's intake of alcohol as a measure of the true average daily intake of alcohol (even if that day's intake were found to be measured without error). Now, in the epidemiological study we chose, for each participant, to measure systolic blood pressure once and then ask them to recall their alcohol intake the previous day. If we now calculate the Pearson product-moment correlation between the two measures we are likely to get a positive value that may be statistically significant (assuming we have a large enough sample) but will not be particularly high (i.e. not far above zero). Suppose, for the sake of argument that we have found a value of this correlation to be 0.20. It should be fairly obvious that as the measures of systolic blood pressure and alcohol get less precise (equivalent for a fixed population to lowering their reliabilities) the correlation will tend to zero. This is attenuation due to measurement error.

Let the observed measurement of blood pressure for the i th participant be Y_i and the corresponding true average blood pressure be τ_i . Similarly, let the measured alcohol intake be X_i with a true average of η_i . We have estimated the correlation between Y and X , ρ_{YX} , when we are really interested in the correlation between the true values, $\rho_{\eta\tau}$. If the errors of measurement for blood pressure are uncorrelated with those for alcohol consumption then it can be shown that the following relationship holds:

$$\rho_{YX} = \rho_{\eta\tau} \sqrt{(\kappa_Y \kappa_X)} \quad (1)$$

Here, κ_Y and κ_X are the reliabilities of the blood pressure and alcohol consumption measurements respectively. It follows that:

$$\rho_{\eta\tau} = \rho_{YX} / \sqrt{(\kappa_Y \kappa_X)} \tag{2}$$

Provided we know the reliabilities for the two measurements, this equation can be used to adjust the observed correlation between Y and X to obtain the required correlation between their true average values. If we know that $\kappa_X = 0.3$ and $\kappa_Y = 0.7$, for example, the required correlation is $0.2 / \sqrt{(0.3 \times 0.7)} = 0.44$.

If, instead of a correlation, the linear regression coefficient for the effect of blood pressure on alcohol consumption were of key interest then:

$$\beta_{YX} = \beta_{\eta\tau} \kappa_X \tag{3}$$

and, again, the required adjustment is straightforward. Equation (3) also holds approximately if we were to use a logistic regression to predict the presence/absence of hypertension.

These calculations are fine as long as we have valid estimates of the reliabilities. However, they are only valid in these very simple situations as described. Epidemiologists almost always wish to adjust their estimates to allow for confounding and some of these confounders are inevitably going to be prone to MEASUREMENT ERROR. Under these circumstances life is considerably more complicated! We cannot even be certain that the estimate of the required parameter will be attenuated, never mind being attenuated in a way described by equation (3). Readers are referred elsewhere to these much more challenging but more realistic situations (Carroll, Ruppert and Stefanski, 1995; Cheng and Van Ness, 1999; Gustafson, 2003). GD

Carroll, R. J., Ruppert, D. and Stefanski, L. A. 1995: *Measurement error in nonlinear models*. London: Chapman & Hall. **Cheng, C.-L. and Van Ness, J. W.** 1999: *Statistical regression with measurement error*. London: Arnold. **Gustafson, P.** 2003: *Measurement error and misclassification in statistics and epidemiology*. London: Chapman & Hall/CRC.

attributable risk As a measure of the public health significance of exposure to a risk factor for disease, the attributable risk provides an estimate of the proportion of diseased subjects that may be attributed to the exposure. It is defined by:

$$\lambda = \Pr\{D\} - \Pr\{D|\bar{E}\}$$

where $\Pr\{D\}$ is the probability that an individual develops disease and E and \bar{E} represent whether an individual is exposed or not exposed to the factor of interest (Levin, 1953). Ideally, one would like to know both $\Pr\{D\}$ and $\Pr\{D|\bar{E}\}$ for the population under study, but for some study designs this is not possible, so if one wishes to use the measure, care is needed to design a study that will provide

as good an estimate as possible, especially when one employs an observational study. Using BAYES THEOREM and rearranging the equation, we can obtain an expression expressed in terms of the relative risk (RR):

$$\lambda = \Pr\{E\}(RR - 1)$$

where $\Pr\{E\}$ is the prevalence of exposure in the population at large. This is a convenient way of expressing the measure of association, because RR is often estimated using alternative study designs, including CASE-CONTROL, COHORT AND CROSS-SECTIONAL STUDIES.

Attributable risk is most easily interpreted when the factor of interest increases risk, i.e. $RR > 1$, and in these cases the possible range of the measure is from 0 to 1. An attributable risk of zero can occur when no individuals in the population are exposed to the factor of interest, or if the factor is not related to risk of disease, $RR = 1$. The measure is not easily interpreted when the exposure is protective, $RR < 1$, so it is generally not used in this case. By redefining the reference group, one can always express the results of a study in a form in which RR is greater than 1, so this is not a serious limitation. In addition, the measure is often expressed as a percent. As RR become large, λ goes to 1, but λ goes to zero either as the proportion exposed, $\Pr\{E\}$, becomes small or as the relative risk, RR , approached the null value of 1. If an entire population is exposed to a particular factor, $\Pr\{E\} = 1$, then the second equation (above) reduces to $\lambda = (RR - 1)/RR$.

The table shows a typical 2×2 table that can be used to display the results from an epidemiological study. In a case-control study, the column totals are generally regarded as being fixed by design and the odds ratio or cross-product ratio is used as a good approximation to the estimate of RR when the disease is rare. In addition, the exposure distribution in the controls, $\Pr\{E\} = \Pr\{E|\bar{D}\}$, is considered to be representative of the exposure distribution in the overall population. Substituting in the sample estimates of these quantities gives rise to what is the maximum likelihood estimate of λ :

$$\hat{\lambda} = \frac{ad - bc}{d(a + c)}$$

attributable risk Results from an epidemiological study with two levels of exposure and disease status

		Disease status		
		D	\bar{D}	Total
Exposed	E	a	b	a + b
	\bar{E}	c	d	c + d
Total		a + c	b + d	N

When setting a confidence interval about the estimate, Walter (1975) suggests using the normal approximation on the log transformation of the complement of the estimate:

$$\text{Var} \left[\log(1 - \hat{\lambda}) \right] = \frac{a}{c(a+c)} + \frac{b}{d(b+d)}$$

Alternatively, Leung and Kupper (1981) have suggested using a logit transformation in which:

$$\text{Var} \left[\log \frac{\hat{\lambda}}{1 - \hat{\lambda}} \right] = \left[\frac{a}{c(a+b)} + \frac{b}{d(b+d)} \right] \left[d(b+d) \right]^2$$

In a cohort study, the row totals in the table are regarded as fixed; therefore such a study does not provide a good internal estimate of the exposure distribution and neither does it provide a good estimate of the unconditional estimate of the probability of disease. In this case, the proportion exposed is usually derived from another study, perhaps an earlier case-control study or a survey of the entire population. A cross-sectional study provides both an estimate of the relative risk and the overall population distribution, so in that sense it is ideal for estimating attributable risk. However, a cross-sectional study suffers in other ways (see CROSS-SECTIONAL STUDIES). Walter (1976) discusses the properties of estimates of attributable risk using these alternative study designs.

Methods for estimating attributable risk for a particular exposure while adjusting for potential confounding factors depends on whether the effect is constant over the levels of the covariates under consideration. When the effect is constant, it can be represented as having a common relative risk over the strata when using a stratified approach, such as the Mantel-Haenszel method, or it can be represented by a main effect only in a model, such as the linear logistic model. In these situations, one can directly use the adjusted estimator or the relative risk, along with an estimate of the exposure distribution in the diseased group in the second equation (above) to obtain an estimate of the adjusted attributable risk (Walter, 1976; Greenland, 1987). However, the assumption that the association can be described without the inclusion of an interaction term is a strong one and it is critical in that a seriously biased estimate can result if it is not true.

An estimate of attributable risk that can be used either in a stratified analysis in which the effect is not homogeneous across strata or in a generalised linear model that includes interaction terms can be expressed as:

$$\lambda = 1 - \sum_{i,j} \frac{P_{ij}}{RR_{ij}}$$

where j represents the levels of the factor(s) being adjusted, i represents the levels of exposure, p_{ij} is the proportion of diseased individuals in (i,j) and RR_{ij} is the relative risk for exposure level i for individuals with level j of the covariates being adjusted (Walter, 1976; Benichou, 1993). TRH

Benichou, J. 1993: Methods of adjustment for estimating the attributable risk in case-control studies: a review. *Statistics in Medicine* 10, 1753–73. **Greenland, S.** 1987: Variance estimators for attributable fraction estimates, consistent in both large strata and sparse data. *Statistics in Medicine* 6, 701–8. **Leung, H. K. and Kupper, L. L.** 1981: Comparison of confidence intervals for attributable risk. *Biometrics* 37, 293–302. **Levin, M. L.** 1953: The occurrence of lung cancer in man. *Acta Unio Internationalis contra Cancrum* 9, 531–41. **Walter, S. D.** 1975: The distribution of Levin's measure of attributable risk. *Biometrika* 62, 371–4. **Walter, S. D.** 1976: The estimation and interpretation of attributable risk in health research. *Biometrics* 32, 829–49.

AUC See AREA UNDER THE CURVE

autocorrelation See CORRELATION

automatic selection procedures These are procedures for identifying a parsimonious model in regression in general and MULTIPLE LINEAR REGRESSION in particular. Such methods are needed because in regression analysis an underfitted model can lead to severely biased estimation and prediction. In contrast, an overfitted model can seriously degrade the efficiency of the resulting parameter estimates and predictions. Consequently a variety of techniques all with the aim of selecting the most important explanatory variables for predicting the response variable and thereby obtaining a parsimonious and effectively predictive model have been developed. Perhaps the three most commonly used methods are *forward selection*, *backward elimination* and a combination of both of these, known as *stepwise regression*.

The forward selection approach begins with an initial model that contains only an intercept and successively adds explanatory variables to the model from the pool of candidate variables until a stage is reached where none of the candidate variables, if added to the current model, would contribute information that is statistically important concerning the expected value of the response.

The backward elimination method begins with an initial model that contains all the explanatory variables being used in the study and then first identifies the single variable that contributes the least information about the expected value of the response; if this is deemed not to be 'significant' then the variable is eliminated from the current model. Successive steps of the method result in a 'final' model from which no further variables can be eliminated without adversely affecting, in a statistical sense, the predicted value of the expected response.

The stepwise regression method combines elements of both forward selection and backward elimination. The initial model considered is one that contains only an intercept. Explanatory variables are then considered for inclusion in the current model, as described previously for forward selection, but now in each step of the procedure variables

included previously are also considered for possible elimination as in the backward method, and they might be removed if the presence of new variables in the model make their contribution to predicting the expected response no longer significant.

In multiple linear regression the criterion used for assessing whether or not a variable should be added to an existing model in forward selection or removed from an existing model in backward elimination is, essentially, the change in the residual sum-of-squares produced by the inclusion or exclusion of the variable. Specifically in forward selection an 'F-statistic', known as the *F-to-enter*, is calculated as:

$$F = \frac{RSS_m - RSS_{m+1}}{RSS_{m+1}/(n-m-2)}$$

where RSS_m and RSS_{m+1} are the residual sums of squares when models with m and $m+1$ explanatory variables have been fitted. The *F-to-enter* is then compared with a preset term; calculated *F*s greater than the preset value lead to the variable under consideration being added to the model. In backward selection a calculated *F* less than a corresponding *F-to-remove* leads to a variable being removed from the current model. In the stepwise procedure variables are entered as with forward selection, but after each addition of a new variable those variables currently in the model are considered for removal by the backward elimination process. (For more details see Petrie and Sabin, 2005.) In other types of regression, for example, LOGISTIC REGRESSION, other criteria are used for judging whether or not a variable should be entered into or removed from the current model. When applying regression techniques to HIGH-DIMENSIONAL DATA more sophisticated variable selection techniques are needed (see, for example, Francois, 2008).

None of the automatic procedures for selecting subsets of variables is foolproof and it is possible for them to be seriously misleading in some circumstances (see Agresti, 1996). That said, at least one can be more confident in a chosen model if all three procedures converge on to the same set of variables, as occurs quite frequently, but not always, in practice. When different subsets of variables are indicated, judgement is necessary to decide on a preferred model, such judgement being based on the desire to create a parsimonious model that is likely to be generalisable, not overly complex as if modelling mere quirks of the particular dataset on which it is based, and yet including important or standard parameters deemed to be of clinical relevance. BSE

[See also ALL SUBSETS REGRESSION]

Agresti, A. 1996: *Introduction to categorical data analysis*. New York: John Wiley & Sons, Inc.. Francois, D. 2008: *High-dimensional data analysis: from optimal metrics to feature selection*. VDM Verlag. Petrie, A. and Sabin, S. 2005: *Medical statistics at a glance*, 2nd edition, Wiley-Blackwell, Chichester.

available case analysis This is an approach to multivariate data containing missing values on a number of variables, in which MEANS, VARIANCES and COVARIANCES (see COVARIANCE MATRIX) are calculated from all available subjects with nonmissing values on the variable (means and variances) or pair of variables (covariances) involved. Although this approach makes use of as much of the observed data as possible, it does have disadvantages. For example, the summary statistics for each variable may be based on different numbers of observations and the calculated variance-covariance matrix may now not be suitable for methods of multivariate analysis such as PRINCIPAL COMPONENTS ANALYSIS and FACTOR ANALYSIS for reasons described in Schafer (1997). BSE

[See also MISSING DATA, MULTIPLE IMPUTATION]

Schafer, J. L. 1997: *Analysis of incomplete multivariate data*. Boca Raton, Florida: Chapman & Hall/CRC.

average age at death This flawed statistic is sometimes used for summarising life expectancy and other aspects of mortality. For example, Andersen (1990) comments on a study that compared average age at death for male symphony orchestra conductors and for the entire US male population and showed that, on average, the conductors lived about 4 years longer. The difference is, however, largely illusory because as age at entry was birth, those in the US male population who died in infancy and childhood were included in the calculation of the average lifespan, whereas only men who survived long enough to become conductors could enter the conductor cohort. The apparent difference in longevity disappeared after accounting for infant and perinatal mortality.

In the other direction, a study in the USA that used average age at death of rock stars (which, on the basis of 321 such deaths, they found to be 36.9 years) to warn of the perils of rock music also got it wrong. It took no account of the rock stars still alive.

Proper analysis of mortality involves the determination of AGE-SPECIFIC RATES for mortality, which requires denominator data on the age distribution of the population (see Colton, 1974). BSE

Andersen, B. 1990: *Methodological errors in medical research*. Oxford: Blackwell Scientific. Colton, T. 1974: *Statistics in medicine*. Boston: Little, Brown and Co.

average treatment effect on the treated (ATT)

See PROPENSITY SCORES

average treatment effect on the untreated (ATU)

See PROPENSITY SCORES