Covariate Imbalance, Adjustment for

Patients in a **clinical trial** tend to vary considerably with respect to clinical and demographic characteristics, some of which may affect their prognosis. It is clearly desirable that the characteristics of the patients in each group are as similar as possible, especially with respect to those characteristics which are prognostic. The enormous strength of randomized controlled trials for drawing **inferences** about treatments is very largely a direct consequence of the use of **randomization** to decide which patients receive each treatment (*see* **Randomized Treatment Assignment**). While randomization eliminates **bias**, it does not guarantee comparable baseline characteristics of the patients in the different treatment groups in a particular trial. Simple randomization is quite likely to yield some differences, especially in small trials. The use of stratified randomization or minimization (*see* **Adaptive and Dynamic Methods of Treatment Assignment**) will reduce such imbalances for selected variables.

Baseline balance is not a requirement. Because of the use of randomization, standard methods of analysis (**estimation** and **hypothesis testing**) will yield valid results regardless of the distribution of baseline variables. Nonetheless, it is often wise to try to avoid imbalance, using the simple design modifications indicated above, and to allow for imbalance if it arises.

Comparison of Baseline Characteristics

An important part of reporting the results of a clinical trial is to describe the patient characteristics for the different treatment groups. As well as characterizing the whole study sample, these data indicate how similar were the groups produced by randomization. In some sense the information is used to determine if the randomization has "worked". Here many investigators behave illogically, by using statistical tests to compare the groups. The aim is probably to attempt to establish that the groups really are comparable, thus strengthening the credibility of the trial. However, the **null hypothesis** for such tests is in essence that the data come from groups which are **random samples** from the same population. Because the treatment groups were indeed random samples, any differences observed between them are necessarily due to chance, and so the use of hypothesis tests is absurd [1]. Yet it is quite common to see groups described as having the "same" characteristics simply because no significant differences were observed, even when, say, there was a notable difference in mean age or the **prevalence** of smoking. A nonsignificant imbalance between groups can be quite important if that **covariate** is highly **prognostic**. Note that such imbalance can work in either direction, masking or overstating the true treatment difference. Significance testing for baseline differences does have one potential use, which is to see if the patients were indeed randomized. However, there is minimal **power** to test this hypothesis. As Senn [17] has noted, a significant imbalance ought really to lead to the conclusion that the trial was not properly randomized – not a conclusion that researchers are likely to draw about their own study. This test can be useful, however, in a **multicenter trial**. Trial coordinators may be able to detect a center which has not adhered to the protocol [8] (*see* **Clinical Trials Protocols**).

Over recent years many authors have examined the practice of testing baseline differences, with unanimous criticism of the practice [1, 3–5, 17]. These papers were mostly published in statistical journals, however, and hypothesis testing of baseline characteristics remains common in reports of trials in medical journals. Baseline testing was found in about 60% of trial reports in recent reviews of general and specialist journals [14]. Such tests are probably quite a recent development – testing was not mentioned in an early paper on baseline imbalance [12]. They are not a requirement of the regulatory bodies [17] (*see* **Drug Approval and Regulation**).

It might be thought that such testing is largely harmless, as it rarely has much impact on how the trial is analyzed and interpreted. However, carrying out one form of analysis conditional on the results of another can distort the results obtained, as described below. (A similar issue arises in other areas of statistics, such as in the analysis of **crossover designs**.) Baseline testing can have other adverse consequences. For example, it has been found that authors selectively report these tests: only 2% of about 1000 tests reported in 206 trial reports in obstetric journals gave results significant at the 5% level [14]. Some of this effect might be due to undisclosed stratification, but it appears that there is a tendency

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a01015

to suppress the results of these tests if the imbalance is significant.

Effect of Imbalance

Imbalance in a patient characteristic will matter only if that characteristic is related to patient outcome, i.e. it is prognostic. In most situations gender will not be prognostic, but age often will be, especially in chronic diseases. Prognostic covariates are most often clinical or biochemical variables, some of which can have major importance. The effect of imbalance for a variable will depend both upon the size of the imbalance (e.g. difference in **means** or proportions) and the strength of the relation between that variable and the outcome.

When randomization leads to baseline imbalance in a prognostic variable, one group will have a poorer prognosis than the other before treatment starts. Thus chance imbalance will lead to a biased estimate of the treatment effect, in either direction according to the direction of the imbalance, when using a simple, unadjusted analysis. Also, a test of significance may yield a significant result when there is no true treatment difference or a nonsignificant result when there is. To take a specific example, Christensen et al. [7] carried out a randomized trial of azathioprine vs. placebo in patients with primary biliary cirrhosis. The unadjusted analysis gave $P = 0.2$ for the treatment comparison. There was some imbalance in serum bilirubin, which is a very strong prognostic variable in such patients. The azathioprine group had higher levels on average and hence a worse prognosis. An adjusted analysis gave $P = 0.02$ for the treatment effect. In practice some imbalance is likely in several prognostic variables, but the overall effect will be much the same as just outlined, especially when, as in this example, one variable is of primary prognostic importance.

Some authors [1, 10, 11] have suggested that imbalance is not so much of a problem for large trials, while others state the opposite [5, 15]. The apparent disagreement arises from the fact that there are several aspects that might be considered – the size of the test of treatment effect, the power of the test, the bias in estimating the treatment effect, and the precision of the estimated treatment effect. Some of these features diminish with increasing sample size, while others apply even to large trials. We certainly cannot rely on large sample size to overcome all of the problems associated with imbalance.

Rationale for Adjusting for Baseline Covariates

There are several reasons why investigators might wish to adjust for baseline characteristics when analyzing the data from a randomized trial, some of which have been mentioned already.

First, as already discussed, the aim of a randomized trial is to compare groups of patients who differ only in that they received different therapies. Imbalance in baseline variables may reduce the credibility of the results, both in the correctness of the randomization procedure and, more importantly, in the validity of the results. Even though such worries may not be well founded, this possibility should be regarded as reasonable grounds for concern. However, the necessary leap in complexity of the methodology (as described below) when adjusting for covariates may be disconcerting to medical readers [9] even though to statisticians it will not cause concern. To some extent transparency is replaced by opaqueness.

Given that chance imbalance in a prognostic variable will lead to some bias in the estimated treatment effect, one of the best reasons for adjusting is to remove this bias. We surely wish to obtain the most reliable estimate of treatment effect. Adjustment will also increase the power to detect a real treatment effect.

Another reason often given for adjusting is to increase the precision with which the treatment effect is estimated. However, while this is the case for normal **regression** models, it will not improve precision in **logistic** [13] or **Cox regression models** [6]. However, in these models, failure to adjust for prognostic variables will lead to underestimation of the treatment effect and hence a reduction in power [6, 13]. The bias associated with not adjusting in non-normal models applies even when there is perfect balance in a prognostic variable.

Methods of Adjusting for Baseline Covariates

The idea behind adjustment for baseline differences is to estimate what the treatment effect would have been if the groups had identical baseline variables, i.e. with

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a01015

identical means for continuous variables and identical frequencies in each group for categorical variables. The generally recommended approach to adjustment is to use regression modeling, with treatment (as a **binary** variable) and prognostic variables included as the explanatory variables. I will follow convention in this context and refer to this approach as **analysis of covariance**, even though these analyses are not all encompassed within the usual idea of that analysis. Analysis of covariance can be used for all types of outcome measure – continuous, binary, and survival times. Its particular strength is that it gives a result that is **unbiased** regardless of the baseline distribution of prognostic variables – i.e. it is conditionally unbiased [16]. In addition, by comparison with an unadjusted analysis, analysis of covariance provides increased precision for the treatment effect (for normal models), an increase in the power of the trial, and a constant conditional size of the test comparing the treatment groups [15].

An alternative is to use a stratified analysis. This approach is more common in epidemiology, where outcomes are usually binary. Pocock [11] gives an example of the use of the **Mantel–Haenszel** test to perform a stratified analysis of a clinical trial. This method is appropriate for categorical covariates, but may not adjust fully for imbalance in continuous covariates [1]. This method may be seen as a special form of analysis of covariance.

There is rather greater difficulty associated with deciding for which covariates to adjust. I consider several possibilities.

Selection Based on Observed Imbalance

The first approach is to focus on the imbalance: two-sample tests, as discussed above, can be used in turn for each prognostic variable to compare the groups at baseline, with no regard to patient outcome. Those which are statistically significant can be used to adjust the treatment effect in a **multiple regression** analysis. While this strategy is very common, its use is unwise. Those variables with significant imbalance may or may not be prognostic, and by including variables conditionally on simple tests, the adjusted analysis is likely to lead to a biased estimate of the treatment effect. Also, as noted above, nonsignificant imbalance may be quite important, even in a normal model.

Selection Based on Relation to Patient Outcome

A second approach is to focus on patient outcome. A multiple regression model can be derived using stepwise selection to see which variables are significant predictors of the outcome, taking no account of baseline balance. While this analysis could be done ignoring treatment or separately within each treatment group, it is most sensible to include all patients and to include in the model an indicator for treatment. This analysis thus yields both the choice of important prognostic variables and the adjusted treatment effect. While far preferable to adjustment based on observed imbalance, this method is not fully satisfactory. Apart from the known overoptimism of regression models based on stepwise selection, adjustment is made for a data-dependent selection of prognostic variables using an arbitrary inclusion rule. We might instead choose those variables which have the largest effect on the estimated treatment effect, either as assessed by the change in the test statistic, as proposed by Canner [5], or the change in the magnitude of the estimated treatment effect (e.g. by 15%). While more reasonable than using ***P* values**, it is unclear here what the criterion should be for deciding which variables have a large enough effect to need adjustment.

An approach proposed by Tukey [19] can be outlined only briefly here. The idea is to minimize the number of regression coefficients without reducing the number of covariates. The outcome variable is regressed on each covariate in turn, with the patients in each group pooled. From each analysis a score is derived from the P value – he suggested scores of 1 to 4 corresponding to $P < 0.05$, $P < 0.01$, $P < 0.001$, and $P < 0.0002$, the scores being signed according to the direction of the effect. The method is easiest to explain with binary covariates each coded as "high" or "low". For a set of covariates, a "composite" is constructed for each patient as a weighted sum of the scores, where the weight is 0 if the variable is low and 1 if high. This composite is then treated as a single covariate to adjust the treatment effect. The weaknesses of this method include the use of the P value as a measure of the strength of the effect, the treatment of all covariates as providing independent information, and the lack of transparency.

Prespecified List of Variables

There are problems associated with all data-derived decisions about which variables to include in an

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a01015

analysis. In particular, the use of significance tests to determine which variables to adjust for is not recommended. It seems far preferable to choose which variables to adjust for without regard to the actual data set to hand.

What criteria should be used to select such variables? Primarily one would wish to consider known important prognostic variables that have not been controlled by the design. It is advisable also to include any variables used for stratification. In addition, in multicenter trials it may be desirable to include centers. The prespecified strategy has the advantage of focusing attention on prognostic factors at the design stage, rather than leaving this issue to be dealt with in an ad hoc manner in the analysis. It means that for some trials the analysis will make adjustment for covariates which are in fact balanced. This will not matter greatly in the case of a **normally distributed** outcome, and is desirable, as noted above, for non-normal outcomes.

Baseline Measurements of the Outcome Variable

In many clinical trials where the object of treatment is to change the value of a continuous measurement (such as blood pressure), it is possible to measure the variable of interest at the start of the trial. The undesirability of baseline imbalance in the variable of primary interest is especially clear. One approach to the analysis of such trials is to analyze change from baseline. This would seem to solve the baseline imbalance problem, but it does not. The change from baseline within each group will usually be highly **correlated** with the baseline values (*see* **Regression to the Mean**), so the difference between the groups in change from baseline will be negatively correlated with the imbalance at the baseline [18]. In other words, while analyzing change from baseline seems to remove the problem associated with baseline imbalance, in fact the chance imbalance will still affect the difference in outcome, but in the opposite direction. In the case where we are seeking to increase lung function, say, and if by chance patients receiving treatment A have higher baseline values than those receiving treatment B, then the analysis of change from baseline will be biased in favor of group B, and vice versa if the imbalance goes the other way. Thus it can be seen that analysis of change from baseline does not deal adequately with baseline imbalance. It is often argued that in such trials change from the baseline is a clinically more relevant outcome measure. Senn [16] has argued strongly against this view. In any case, one can use analysis of covariance to adjust change from baseline for baseline values, with exactly equivalent answers, so the debate is irrelevant if analysis of covariance is used [16, 18] (*see* **Baseline Adjustment in Longitudinal Studies**).

Such trials may cause a further error. It is quite common to see authors report separate tests to assess whether each group has changed from the baseline. The resulting P values are compared and a difference claimed when one P value is significant and the other is not. This is not a valid form of statistical inference, and is likely to mislead [2, 15].

Comments

The main issues here are the proper analysis of randomized trials, and the distinction between substantive and **exploratory analyses**. A clear recommendation may be made for the analysis of trials. Ideally, a prespecified strategy should be developed as part of the protocol in which either no adjustment will be made for baseline variables or adjustment will be made for nominated variables using analysis of covariance.

Good statistical practice requires investigators to prespecify in the study protocol their intentions with regard to sample size (*see* **Sample Size Determination**), primary (and subsidiary) endpoints (*see* **Outcome Measures in Clinical Trials**), subgroup analyses (*see* **Treatment-covariate Interaction**), and so on. It is no different to suggest that the analysis strategy should also be prespecified, in particular intentions regarding adjusted analyses. It is usually known in advance which are the variables that are most prognostic of patient outcome. Whether or not these are used as stratifying variables, the trial protocol should specify which ones will be adjusted for in the analysis, and that this adjustment will not be conditional on the distribution of those variables across the treatment groups [4, 5, 15]. It would not be acceptable to specify in the protocol that adjustment would be made for any variables showing statistically significant imbalance; this would not circumvent the problems described above.

The protocol should also specify which will be the primary analysis. My view is that this should usually

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a01015

be the adjusted analysis, otherwise there is little point in performing it. Sometimes, however, the adjusted analysis may be performed in order to strengthen belief in the results of the unadjusted analysis. Here it becomes unclear how similar the results need to be for the unadjusted analysis to be confirmed. It is not desirable for the choice of primary analysis to be conditional on the results.

It may not be as simple as just suggested to identify the "most prognostic variables". Clinicians may argue that all information being collected is potentially prognostic, which is why it is being collected. It may prove difficult to persuade them to identify in advance which variables will and which will not be adjusted for in the analysis, and harder still to get them to comply with this strategy when imbalance is seen within the latter group. Despite the wide recommendation of this general strategy, it is not common to see published studies reporting this as the basis for their chosen analysis. Partly, though, this may be because balance has been achieved through the design.

In practice, imbalance may arise when the possible need for adjustment has not been anticipated. What should the researchers do? They might choose to ignore the imbalance; as noted, this would be entirely proper. The difficulty then is one of credibility. Readers of their paper (including reviewers and editors) may question whether the observed finding has been influenced by the unequal distribution of one or more baseline covariates. It is still possible, and arguably advisable, to carry out an adjusted analysis, but now with the explicit acknowledgment that this is an exploratory rather than definitive analysis, and that the unadjusted analysis should be taken as the primary one. Obviously, if the simple and adjusted analyses yield substantially the same result, then there is no difficulty of interpretation. This will usually be the case. However, if the results of the two analyses differ, then there is a real problem. The existence of such a discrepancy must cast some doubt on the veracity of the overall (unadjusted) result. The situation is similar to the difficulties of interpretation that arise with unplanned subgroup comparisons. One suggestion in such circumstances is to try to mimic what would have been done if the problem *had* been anticipated, namely to adjust not for variables that are observed to be unbalanced, but for all variables that would have been identified in advance as prognostic. An independent source could be used to identify such variables. Alternatively, the trial data could be used to determine which variables are prognostic. This strategy too could be prespecified in the study protocol. Because this analysis would be performed conditionally on the observed imbalance, it does not remove bias and thus cannot be considered fully satisfactory.

Finally, I have assumed implicitly that the treatment effect is the same on average regardless of the values of the covariates. It may be desirable to examine whether there are any **treatment–covariate interactions**; here, too, prespecification of intentions is strongly advisable.

References

[1] Altman, D.G. (1985). Comparability of randomised groups, *Statistician* **34**, 125–136.

[2] Altman, D.G. & Doré, C.J. (1990). Randomisation and baseline comparisons in clinical trials, *Lancet* **335**, 149–153.

[3] Beach, M.L. & Meier, P. (1989). Choosing covariates in the analysis of clinical trials, *Controlled Clinical Trials* **10**, 161S–175S.

[4] Begg, C.B. (1990). Significance tests of covariate imbalance in clinical trials, *Controlled Clinical Trials* **11**, 223–225.

[5] Canner, P. (1991). Covariate adjustment of treatment effects in clinical trials, *Controlled Clinical Trials* **12**, 359–366.

[6] Chastang, C., Byar, D. & Piantadosi, S. (1988). A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models, *Statistics in Medicine* **7**, 1243–1255.

[7] Christensen, E., Neuberger, J., Crowe, J., Altman, D.G., Popper, H., Portmann, B., Doniach, D., Ranek, L., Tygstrup, N. & Williams R. (1985). Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis: final results of an international trial, *Gastroenterology* **89**, 1084–1091.

[8] Collins, R., Gray, R., Godwin, J. & Peto, R. (1987). Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews, *Statistics in Medicine* **6**, 245–250.

[9] Greenberg, E.R., Baron, J.A. & Colton, T. (1983). Reporting the results of a clinical trial, in *Clinical Trials: Issues and Approaches*, S.H. Shapiro & T.A. Louis, eds. Marcel Dekker, New York, pp. 191–204.

[10] Grizzle, J.E. (1982). A note on stratifying versus complete random assignment in clinical trials, *Controlled Clinical Trials* **3**, 365–368.

[11] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. Wiley, Chichester, pp. 211–221.

[12] Radhakrishna, S. & Sutherland, I. (1962). The chance occurrence of substantial initial differences between

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a01015

groups in studies based on random allocation, *Applied Statistics* **11**, 47–54.
[13] Robinson, L.D. & Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression models, *International Statistical Review* **58**, 227–240.
[14] Schulz, K.F., Chalmers, I., Grimes, D.A., Altman, D.G. & Doré, C.J. (1995). The methodologic quality of randomization as assessed from reports of trials in specialist and general medical journals, *Online Journal of Current Clinical Trials* **4**, Doc. No. 197.
[15] Senn S. (1989). Covariate imbalance and random allocation in clinical trials, *Statistics in Medicine* **8**, 467–75.
[16] Senn, S. (1991). Baseline comparisons in randomized clinical trials, *Statistics in Medicine* **10**, 1157–1159.
[17] Senn, S. (1995). Base logic: tests of baseline balance in randomized clinical trials, *Clinical Research and Regulatory Affairs* **12**, 171–182.
[18] Senn, S. (1997). *Statistical Issues in Drug Development*. Wiley, Chichester, pp. 95–109.
[19] Tukey, J.W. (1991). Use of many covariates in clinical trials, *International Statistical Review* **59**, 123–137.

(*See also* **Variable Selection**)

DOUGLAS G. ALTMAN

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a01015