12.4: Variability and quality control of outcome measures

4.1 Reproducibility

The extent to which different observers will make the same diagnoses or assessments on a participant and to which observers are consistent in their classifications between participants may have an important influence on the results of a trial. Clearly, it is desirable to choose outcome measures for which there is substantial reproducibility and agreement among observers, with respect to the classification of participants in the trial.

For objective outcome measures, variations between observers, or by the same observer at different times, may be small and unlikely to influence the results of a study. For outcome measures requiring some degree of subjective assessment, however, such variations may be substantial. The likely degree of such variations will influence the choice of outcome measures, as it will be preferable to select those measures that have the smallest inter- and intra-observer variations, yet still give valid measures of the impact of the intervention.

Variation among observers is often much greater than expected, for example, in the reading of a chest X-ray to assess whether there is evidence of pneumonia. If a study involves several observers, pilot studies should be conducted, in order to measure the extent of the variation and then to seek to standardize the assessment methods to minimize the variation. With suitable training, it is usually possible to reduce the variation between observers substantially.

For some outcomes, independent assessment by two observers should be routine, with a third being called in to resolve disagreements. It may be costly to screen the whole trial population in this way, but a common approach is to have all suspected cases of the disease of interest examined by a second observer, mixed in with a sample of those not thought to have the disease. Sometimes, it is possible to have the observer examine the same individual twice, but these examinations may not be independent, unless the survey is large and the observer does not remember the result of the
first assessment.

It is important to make every effort to reduce variability to the maximum extent possible. Having done so, however, it is also critical to know the extent of the remaining ‘irreducible’ variability for purposes of analysis. The purpose of trials is usually to demonstrate the effect of an intervention or to compare differences between interventions. Knowledge of the inherent variability in diagnostic procedures is essential for this demonstration, and the best way of assessing this is through replicate measures. It is especially important to take account of between-observer differences when communities are the units of randomization in a field trial. Differences between observers may produce biases if different observers are used in different communities. In such situations, it is better to organize the fieldwork so that the workload within each community is split among different observers and differences between the observers are not confounded with the effect of the intervention.

4.2 Sensitivity and specificity

The choice of an appropriate definition of a ‘case’ in a field trial will be influenced by the sensitivity and specificity associated with the diagnostic criteria. Sensitivity is defined as the proportion of true cases that are classified as cases in the study. Specificity is the proportion of non-cases that are classified as non-cases in the study. A low sensitivity is associated with a reduction in the measured incidence of the disease. This decreases the likelihood of observing a significant difference between two groups in a trial of a given size. In statistical terms, it reduces the power of the study (see Chapter 5, Section 2.2). If the incidence of the disease in both the intervention group and the comparison group will be affected proportionately in the same way, as is often the case, it does not bias the estimate of the relative disease incidence in the two groups, though the absolute magnitude of the difference will be less than the true difference. Thus, in the context of a vaccine trial, because protective efficacy is assessed, in terms of relative differences in incidence between groups, the estimate of protective efficacy will not be biased, but the confidence limits on the estimate will be wider than they would be using a more sensitive case definition. In theory, the reduction in power associated with low sensitivity can be compensated for by increasing the trial size.

In general, a low specificity of diagnosis is a more serious problem than a low sensitivity in intervention trials. A low specificity results in the disease incidence rates being estimated to be higher than they really are, as some participants without the disease under study are classified incorrectly as cases. Generally, the levels of inflation in the rates will be similar, in absolute terms, in the intervention and comparison groups, and thus the ratio of the measured rates in the two groups will be less than the true ratio, though the difference in the rates should be unbiased. Thus, in vaccine trials, for example, the vaccine efficacy estimate will be biased towards zero, though the absolute difference in the rates between the intervention and control groups will not be biased (unless there is also poor sensitivity). Increasing the trial size will not compensate for the bias in the estimate of vaccine efficacy.

In algebraic terms, suppose the true disease rates are \( r_1 \) and \( r_2 \) in the two groups under study, the true relative rate \( R \) is \( r_1/r_2 \), and the true difference in disease rates \( D \) is \( r_1 - r_2 \). If sensitivity is less than 100% (but specificity is 100%), and only a proportion \( k \) of all cases are correctly diagnosed, the measured disease rates in the two groups will be \( kr_1 \) and \( kr_2 \); the measured relative rate will be \( kr_1/(kr_2) = R \); and the measured difference in disease rates will be \( kr_1 - kr_2 = k(r_1 - r_2) = kD \) (which will be less than \( D \)). If specificity is less than 100% (but sensitivity is 100%), and the rate of false diagnoses is \( s \), the measured rates in the two groups will be \( (r_1 + s) \) and \( (r_2 + s) \); the measured relative rate will be \( (r_1 + s)/(r_2 + s) \) (which will be less than \( R \)); and the measured difference in disease rates will be \( (r_1 + s) - (r_2 + s) = D \).
To measure the sensitivity and specificity of the diagnostic procedures used in a trial, it is necessary to have a ‘gold standard’ for diagnosis (i.e. it is necessary to have a diagnostic procedure that determines who really is a case and who is not). Sometimes, this is not possible, and, even if definitive diagnostic procedures exist, it may be necessary to use imperfect procedures in a field trial for reasons of cost or logistics. In this situation, if an assessment is made of sensitivity and specificity, it is possible to evaluate the consequences for the results of a field trial, and possible even to correct for biases in efficacy estimates due to the use of a non-specific diagnostic test. Unfortunately, in many situations, there is no ‘gold standard’, and so the sensitivity and specificity of the diagnostic methods used remain uncertain. For example, there is no universally agreed definition of a case of clinical malaria. Most would agree that the presence of parasites in the blood is necessary (unless a potential case has taken treatment before presenting to the study clinic), and many would agree that the presence of fever associated with parasitaemia increases the likelihood of the disease being clinical malaria, but it is also possible that the fever is due to other causes, rather than the parasitaemia being the cause of the fever.

The bias induced by a low specificity of diagnosis is most severe for diseases that have a low incidence. A good example of this is provided by leprosy, which is both difficult to diagnose (in the early stages) and also of low incidence. Consider a vaccine trial in which the true disease incidence in the unvaccinated group is ten per thousand over the period of the trial, and the true efficacy of a new vaccine against leprosy is 50%, i.e. the true disease incidence in the vaccinated is five per thousand over the period of the trial. If the sensitivity of the diagnostic test used for cases is 90%, but the specificity is 100%, the observed disease incidences would be $10 \times 0.9 = 9.0$ and $5 \times 0.9 = 4.5$ per thousand, respectively. Thus, the estimate of vaccine efficacy is correct (50%). The power of the study is reduced, however. To achieve the power that would be associated with a ‘perfect’ test, the trial size would have to be increased by about 11%.

On the other hand, if the specificity of the diagnostic test is as high as 99% and the sensitivity is 100%, the observed disease incidences would be ten true cases + $(990 \times 0.01 = 9.9)$ false cases = 19.9 per thousand in the unvaccinated group, and five true cases + $(995 \times 0.01 = 9.95) = 14.95$ per thousand in the vaccinated group. Thus, even with a test with 99% specificity, the estimate of vaccine efficacy is reduced from the true value of 50% to 25%. If the specificity of the test were 90%, the expected estimate of vaccine efficacy would be only 4%.

In vaccine trials, the sensitivity and specificity of the diagnostic test are of consequence in different ways at different times in the trial. When individuals are screened for entry to the trial, it is important that the test used should be highly sensitive, even if it is not very specific, as substantial bias may be introduced if undiagnosed ‘cases’ are included in the trial and included in the vaccinated or unvaccinated groups. If the vaccine has no effect on the progression of their disease and they are detected as cases later in the trial, a false low estimate of efficacy will result. Thus, individuals whose diagnosis is ‘doubtful’ at entry to the trial should be excluded from the trial. Conversely, once individuals have been screened for entry into the trial and they are being followed for the development of disease, a highly specific test is required to avoid the bias illustrated in the preceding paragraph.

In situations where there may be no clear-cut definitions of a case (for example, early leprosy or childhood TB), studies of intra- and inter-observer variation may be undertaken, using various definitions of the disease. The definition that shows the least disagreement between observers and gives maximum consistency within each observer may be the appropriate one to use in a trial, but the investigator should be aware of the potential for bias if the specificity of the diagnostic procedure is less than 100%.
4.3 Bias

The most powerful way to minimize bias in the assessment of the impact of an intervention is through the conduct of a double-blind randomized trial. If these two aspects are built into a trial, an effect of an intervention is not likely to be observed if there is no true effect. However, as pointed out in Section 4.2, if the specificity of the diagnosis for the outcome of interest is poor, the estimate of the efficacy of an intervention, measured in relative terms, may be biased towards zero, even in a properly randomized double-blind investigation.

It is highly desirable that the person making diagnoses in a trial is ignorant of which intervention the suspected cases have received. If the diagnosis is based on laboratory tests or X-ray examinations, blindness should be easy to preserve. In some circumstances, it may be possible to determine from the results of a laboratory test which intervention an individual has received, as the test may be measuring some intermediate effect between the intervention and the outcome of prime interest (for example, an antibody response to a vaccine). In such cases, those making diagnoses in the field should not be given access to the laboratory results. For example, in placebo-controlled studies of praziquantel against schistosomiasis in communities where the infection is common, those who had received the active drug would be easily detected by a rapid reduction in egg counts in stool or urine samples following treatment. If the outcome of main interest is morbidity from the disease, then the egg count information should be kept from those making the assessment of morbidity. It would generally be inappropriate to use measures of antibody level to make diagnoses of disease following vaccination, if the vaccination itself induced antibodies indistinguishable from those being measured. Similarly, tuberculin testing should not be part of diagnostic procedures for TB in studies of the efficacy of BCG vaccination, as the vaccine alters the response to the test.

If the diagnosis of disease is based on a clinical examination, it may be necessary to take special precautions to preserve blindness. An example is given in Chapter 11, Section 4, with respect to a BCG trial against leprosy, in which all participants had the upper arm area, where BCG or placebo was injected, covered during the clinical examination, since BCG leads to a permanent scar. Even if the participants know which intervention they had, it is important to try to keep this knowledge from the person making any diagnoses. Thus, participants might be instructed not to discuss the intervention with the examiner, and the examiner would be similarly restricted. Such a procedure is obviously not fail-safe, but great efforts should be made to preserve blindness, if at all possible, especially if the diagnosis is made on subjective criteria.

If randomization in a trial is by community, rather than by individuals, it may be especially difficult to keep examiners ignorant of the intervention an individual received. Sometimes, ways can be found of doing this, for example, by conducting surveys for disease by bringing all participants to a clinic outside the trial communities. If communities are randomized to receive an improved water supply or not, one outcome measure of interest might be the incidence of scabies infection. It may be difficult to avoid the possibility of the diagnoses of scabies being influenced by the observer’s knowledge of whether or not the participant was in a village with an improved water supply. In such a case, it may be best to seek other measures of impact, based upon objective criteria or laboratory measures, or to take photographs of the relevant body parts and have these assessed objectively and ‘blind’ to intervention group.

4.4 The Hawthorne effect

Trials that require active home visits by study personnel during the surveillance period to evaluate the effect of an...
intervention may be affected by an indirect effect of the home visits on the study objective, even when not intended. The presence of a study member in a subject’s home may have a positive effect on the health status of the subject, since it may, for example, stimulate better health behaviour of the subject or improve hygiene practices in the house or better health care utilization. In studies with such effects, rates of illnesses or of severe illness may be reduced in both study arms—an indirect effect known as the ‘Hawthorne effect’ (named after a study in the 1930s in the USA at the Hawthorne Works, in which it was documented that worker behaviour changed as a consequence of them being observed). This effect reduces the power of the study and may make it inconclusive. There is no easy way to control for it, so, if such a Hawthorne effect is expected in a field trial, the sample size may need to be increased to maintain statistical power.

4.5 Quality control issues

The sensitivity and specificity of the diagnostic procedures employed in a trial should be monitored for the duration of the trial, as they may change as the study progresses. Such changes may be for the worse or for the better. With experience, diagnostic skills may improve, but also, as time passes, the staff may become bored and take less care. It is important that the field staff are aware that their performance is being continuously monitored. If this is done, then anyone who goes ‘off the rails’ can be steered back or removed from the study, before much harm is done. Such monitoring is important for both field and laboratory staff.

The methods used to monitor the quality of diagnostic procedures may include the re-examination of a sample of cases by a supervisor or a more highly trained investigator

and, for the laboratory, may be done by sending a sample of specimens to a reference laboratory and by passing some specimens through the laboratory in duplicate, in a blinded fashion, to determine if the differences between results on the same specimen are within acceptable limits (see Chapter 17, Section 5).

If the disease under study is relatively rare, it may be difficult to measure sensitivity based on small numbers of individuals being examined twice. While it will be possible to check if specificity is poor (a high proportion of those classified as cases are wrongly diagnosed), checks on sensitivity may involve the examination of thousands of individuals twice to determine if cases are being missed. Fortunately, in most trials, specificity is of more critical importance than sensitivity, although the relative importance can change as the survey goes on, as discussed in Section 4.2.