NCBI Bookshelf. A service of the National Library of Medicine, National Institutes of Health.

Smith PG, Morrow RH, Ross DA, editors. Field Trials of Health Interventions: A Toolbox. 3rd edition. Oxford (UK): OUP Oxford; 2015 Jun 1.

Chapter 12 Outcome measures and case definition

1. Introduction to outcome measures and case definition

Field trials of health interventions are designed to assess the impact of one or more interventions on the incidence, duration, or severity of specified diseases, or on intermediate variables or risk factors considered to be closely related to these measures of disease (for example, hygiene behaviours for diarrhoeal diseases, reduction in density of parasite vector, reduction of indoor air pollutants for pneumonia, or reduction of salt intake for hypertension). The measures chosen to assess the impact of the interventions are called the *outcome* measures in the trial (or the trial *endpoints).* Such measures should be defined at the time the trial is designed and should be specified in detail in the study protocol. The outcomes should be compared between those in the different intervention groups and should be measured in a consistent way during the course of the trial in the different groups. Clear definitions are also necessary, so that the measures can be replicated in other trials and meaningful comparisons made between trials. Failure to pay sufficient attention to the precise definition of the primary outcome measures at the start of a trial may lead to confusion in interpreting the results or can even invalidate them.

As discussed in Chapter 4, Section 5, several different outcome measures may be employed in a trial. It is important to decide which is of most interest (primary outcome), as this has major design implications, particularly in terms of the study size and duration. Trials may have other outcomes (secondary or tertiary) that may be important to measure, although they will generally not determine the size of the trial. In Table 12.1, there are some examples of primary and secondary outcomes for trials of different interventions.

In this chapter, different types of outcome measures are reviewed in Section 2, and factors influencing the selection of these are discussed in Section 3. The importance of standardizing measurements between different observers is stressed in Section 4.1, and there is a discussion of how the results of a trial may be influenced by poor sensitivity or specificity in the outcome measures in Section 4.2. Finally, ways of avoiding bias and maintaining quality control (QC) in case ascertainment methods are reviewed in Sections 4.3 and Sections 4.5.

2. Types of outcome measures

2.1. Primary, secondary, tertiary

2.1.1. Primary outcomes

Primary outcomes are the most important outcomes of the study, the ones that determine its design and the study size. They represent the main reason the trial is being conducted. Normally, a trial has only one primary outcome, so, for each main question in the development of a new drug, vaccine, or intervention, one specific trial is usually conducted. However, more than one primary outcome may be selected in some trials, provided the design and sample size allow it and if measuring them in the study does not substantially add to the cost or complicate the design or conduct of the trial. For example, Phase I or II clinical trials usually have several primary outcomes (such as the safety of a new drug or vaccine, evaluated through a series of clinical outcomes, as well as the immunogenicity of the vaccine or pharmacodynamics of the drug). Phase III or IV trials have fewer primary outcomes, and often only one. Primary outcomes need careful definition prior to the start of the trial (indicator, instrument to be used, measurement to be taken, values which will be considered as a positive or negative result, which laboratory will be used, etc.); these should be agreed upon among investigators, sponsors, and any regulatory agencies overseeing the trial.

2.1.2. Secondary and tertiary outcomes

Trials often have additional important outcomes, but these are not usually used to determine the trial design and sample size. They are included as secondary or tertiary outcomes to be measured in the trial. These outcomes may not be statistically conclusive, since the trial may not have been designed with the power to evaluate them, but they can be very useful to generate further hypotheses and guide future trials. Because of their importance in justifying future studies, these additional outcomes also need careful definition and measurement and should be fully specified in the protocol, since extra resources often are needed to measure and evaluate them.

2.1.3. Other variables which are not study outcomes

Often, trials have other variables measured in the study not directly related to the study outcomes. Variables, such as age, gender, educational or socio-economic level, and nutritional status, may be used to evaluate potential effect modifiers or confounders to the study outcomes. These variables also need to be defined and considered at the beginning of the study, so they may be included in any pilot investigations.

2.2. Clinical case definitions

2.2.1. Physician-based case definitions

In some trials, outcomes are based on a clinical diagnosis by a physician, without any type of laboratory confirmation. For example, pneumonia may be diagnosed by auscultation in a trial evaluating the impact of an intervention designed to reduce indoor air pollution. This type of outcome is subjective, and interpretation may vary among doctors, and even among experienced specialists. Nevertheless, in many clinical trials, physician-based clinical diagnosis determines the main outcome of the study, since no alternatives exist. For many diseases, standardized criteria for defining a 'case' have been established by experts. The International Classification of Diseases (World Health Organization, 2010; see also <http://www.who.int/classifications/icd/en>), which is revised about every 10 years, provides a basis for coding all diseases in a systematic way and is widely used for clinical and epidemiological research.

If standardized criteria for a 'case definition' have not been developed for the disease under study, a suitable definition should be established before the trial starts. For infectious diseases, there is often the need to distinguish between infection and disease, since clinical manifestations of infections may vary widely, from subclinical to overwhelming disease. For many trials, the main outcome of public health interest may be those infections that are severe or fatal. Careful definitions of these types of clinical categories are important, and, if available, the criteria used in other studies should be used to facilitate comparability across studies. The physicians charged with making diagnoses in the trial should discuss and agree the criteria they will be using to make a diagnosis and should compare their diagnoses on a range of patients prior to the start of the trial and at periodic intervals throughout the trial (see Section 2.2.5). Cases may also be classified as suspected, probable, or definite, using clinical and/or laboratory criteria.

In some populations, the conduct of a clinical examination may be problematic. Physical examinations are virtually always highly personal and may raise sensitive issues concerning individual dignity. In those populations when privacy is required, a third person in the examination room is often important, both to reassure the patient and to provide protection against possible charges of misconduct. In the case of children, the mother's presence should normally be requested; for the examination of women, a nurse and an appropriate family member may be needed, even when the examiner is a woman. If there are local codes of behaviour that cover such circumstances, these must be adhered to.

2.2.2. Laboratory-based case definitions, including any diagnostic procedure

Commonly, a clinically defined study outcome involves the combination of a clinical assessment with the support of a confirmatory laboratory, or other diagnostic, procedure. For example, the clinical diagnosis of malaria may be supported by a positive identification of the parasite in the blood, or the diagnosis of dengue fever in a subject with 48 hours of elevated temperature with a positive immunoglobulin M or viral antigen present in the blood, as detected by polymerase chain reaction, or the clinical diagnosis of pneumonia with a confirmatory chest X-ray. All these diagnostic procedures need careful definition, including the technique, machine, or equipment to be used, reference values considered normal for the study population, and the level at which they will be considered abnormal. It is important to describe, in the protocol, how the test or procedure will be conducted and whether a reference laboratory will be used to validate the site laboratory or procedure—also, how procedures used by laboratory personnel to interpret results will be standardized and how monitoring for QC will be done. Some diagnostic results are also affected by subjectivity such as reading the results of a chest X-ray. In such cases, protocols have been developed to try to standardize the diagnosis, such as establishing defined criteria for each type of pathology in advance, having two independent, blinded radiologists read all X-ray films, with a third radiologist reading all films where there were disagreements, with their result used as the tiebreaker. Similar procedures have been developed to read blood smears for malaria. All these options have important consequences on the trial logistics and cost, so careful consideration

needs to be given to them when designing the trial and selecting its study outcomes. Issues concerning laboratory tests of relevance to diagnosis in field trials are outlined in Chapter 17.

2.2.3. Lay worker-based case definitions

Some trials use lay workers (fieldworkers) to measure a study outcome. Examples of such trials are diarrhoeal diseases where prevalent diarrhoea might be defined as three or more liquid or semi-liquid stools passed in a 24-hour period, as reported by the mother or the child's caretaker to a fieldworker, or hygiene behaviours observed by fieldworkers in spot household checks during a hand-washing intervention trial. These types of outcomes are usually captured in questionnaires or study forms. Interviewing techniques and questionnaire design are discussed in Chapter 14.

Fieldworkers may also measure a clinical indicator such as the body temperature or respiratory rate. Because of the high cost of using physicians, in many trials, lay workers or paramedical workers are trained to assess clinical signs and symptoms. When using lay workers or professional fieldworkers, such as nutritionists, auxiliary nurses, or nurse technicians, it is essential to train them and standardize the methods they use, in order to assure uniform implementation of these procedures in the field throughout the study, with good supervision and QC procedures.

2.2.4. Case definitions using secondary data sources

In some trials, such as in phase IV trials, existing surveillance systems may be used to define a study outcome. These secondary data sources, in which trial outcomes are not measured directly by study staff, will have the limitations intrinsic to the quality of the existing surveillance system. Examples of such study outcomes are post-marketing passive surveillance of vaccine or drug-related SAEs, such as hospitalizations of any type, after the introduction of the intervention into general use. They could also be used to evaluate the efficacy of a new vaccine or intervention on an important outcome which, for reasons of cost or ethics, could not be measured in a phase III trial such as the impact of a new vaccine on mortality.

2.2.5. Standardization

All study outcomes to be used in a clinical trial need to be properly standardized. When an outcome requires physicians, other professionals, or lay workers to measure it, standardization usually requires predefined exercises, with the use of an expert to act as the 'standard' against which the group is compared, defining differences which will be considered acceptable as part of the precision of the study. These standardization exercises could be done with real patients or mock subjects who may be trained actors. The use of videos showing different types of patients, which all participants evaluate independently, is a very useful exercise to help standardize them against the 'standard' observer. Standardization of this sort is not easy; it requires resources, time, and, in many cases, patients or volunteers willing to be examined by multiple persons. Ideally, the same set of samples, films, blood smears, subjects, or videos would be evaluated again by the same individual in a random order, under code, to allow the calculation of intra-observer reproducibility. All these procedures need to be carefully described in operating manuals and recorded, so they can be reviewed by investigators, collaborators, or regulatory agencies. In studies that last for several years, it is important to re-standardize observers every 6 to 12 months or if any observer needs to be replaced, to assure that the quality of the study is maintained.

2.2.6. Inclusion and exclusion criteria

An important component of an outcome definition is the description of the inclusion and exclusion criteria for the subjects to be evaluated in the trial. Ideally, the trial results should be able to be generalized to the whole population in which the intervention will be used. Under ideal circumstances, nobody should be excluded from the trial. However, for ethical, logistic, or analytical reasons, most trials establish stringent inclusion and exclusion criteria to exclude certain persons from participation. These criteria could be established on the basis of factors such as age, gender, literacy, being healthy or not, not affected by chronic diseases different from the study outcome, or not affected by other conditions such as abnormal baseline laboratory results. All these criteria need careful evaluation and discussion not only within the research team and the sponsor of the trial, but also with the ethics committees, the regulatory agencies overseeing the trial, and the communities in which the trial will take place, to assure that the trial results can be generalized to the intended population. It is common practice to exclude persons who are very sick from a trial (unless, of course, the trial intervention is directed at such persons). This is done because early deaths, or other SAEs

in such persons, may occur independently of the trial intervention but may complicate interpretation of the effects of the intervention.

Signing a written informed consent form is now a standard inclusion criterion in most clinical trials (see Chapter 6). However, such a requirement will select a subgroup of the population who accept to sign such a form and participate in the study, generating a potential selection bias. To measure how strong that bias may be, it is important to register all eligible subjects who were considered as potential participants in the trial, indicating the reasons for refusal for those who did not enter into the trial.

2.3. Death and verbal autopsies

Preventing deaths (or severe disabilities) is one of the most important public health outcomes of any type of treatment or preventive intervention. It is the most important outcome in driving disease control policies and the introduction of new interventions or treatments into the population, once they have been found to be safe and effective. These types of outcomes have the heaviest weight in terms of disability-adjusted life-years (DALYs), when undertaking cost-effectiveness analyses of new drugs or interventions (see Chapter 19). Therefore, trials designed to evaluate these outcomes are very important. But, for many reasons, they may be difficult and costly to conduct, and, in many cases, they may not be feasible or ethical to do. Counting deaths in the conduct of a trial is a very sensitive issue, particularly in developing countries with poor health systems. It may create moral issues or generate political tension that may stop the trial. Therefore, few trials are done with these important outcomes, despite their major importance. However, those trials that are done with this endpoint and which demonstrate that an intervention significantly reduces mortality are most likely to influence a policy decision on a more widespread introduction of the intervention.

When deaths or severe disability are chosen as study outcomes, several problems emerge, depending on the setting where the study is conducted. In many LMICs, the quality of vital registration systems is poor or they are non-existent, precluding their use. Therefore, methods are needed to identify deaths, as well as to establish causes of death. In LMICs, the most commonly used method to ascertain causes of death are ‘verbal autopsies’. A verbal autopsy is a structured interview, conducted with the relatives of the deceased person, with the intention to reconstruct the series of events that led to the death (or severe complication or disability). Standard verbal autopsy questionnaires have been developed (World Health Organization, 2012). Such ‘autopsies’ should be conducted neither too soon after the death (to avoid asking questions when relatives are still very upset by the death) nor too long after the death (to avoid recall bias). This interview is then analysed in a standardized way, either by physicians or using a computer algorithm, to classify the likely cause of the death, following a predefined set of criteria (Lopez et al., 2011).

The reliability of verbal autopsy methods varies according to the cause of death, as some causes of death may be confused because signs and symptoms in the illness leading up to death may be similar. The usefulness of verbal autopsies is also dependent on the culture of the population under surveillance. It is essential to pilot-test the (translated) questionnaire to assure that appropriate local words are used to ascertain signs or symptoms of the causes of death.

In many populations, there could be a wide range of reasons why deaths may not be reported, and therefore special care should be taken to ensure that ascertainment is as complete as possible. This becomes crucial when the study outcome is death in the perinatal period, since an important proportion of live births that die in the minutes or hours after birth could be either missed or wrongly reported as stillbirths. In some trials, members of the study community may be hired as local informants to report any deaths. Other techniques include enumerating all members in a community and checking for the absence of any of them in frequently conducted cross-sectional surveys. Special attention should be paid to households for which all members are absent during one of these follow-up surveys, because the death of an adult may lead to dissolution of a household or migration of household members. Enquiries should be made with neighbours in such circumstances. Training and standardization of interviewers are essential. The frequency of surveillance will be a critical decision in designing trials with mortality outcomes, since a long recall period (such as 1 year) may miss deaths, particularly of children or infants; but each additional surveillance round will be expensive.

2.4. Non-clinical case definitions

Non-clinical case definitions can also be used in trials such as quality of life in trials of the use of chemotherapy for advanced cancer, antibiotic use in children in settings where they are available without prescription, satisfaction of users of a health service, and economic outcomes (costs) which are discussed in Chapter 19. They also may include

outcomes that come directly from patients about how they feel or function in relation to a health condition and its therapy (so called *patient-reported outcomes*), without interpretation by health care professionals or anyone else. For these case definitions, instruments that have been developed previously or that are created especially for the trial need to be validated, in order to have valid and comparable results.

2.5. Proxy measurements as study outcomes

Some trials may select outcome measures that are associated with the outcome of interest such as reported risky sexual behaviour, which are either easier to measure, cheaper, or more socially acceptable. Those outcomes are called 'proxy' measurements of the outcome of interest. Such measures, however, may be subject to invalidity and bias (for example, misreporting, differential degrees of desirability bias between trial arms).

2.5.1. Behavioural changes

A behaviour thought to be critical to reduce the disease of interest might be selected as a study outcome. For example, in a study to investigate the effectiveness of a health education campaign to promote the use of latrines, where the ultimate objective was to reduce diarrhoeal disease, the frequency of use of latrines might be measured. Sometimes, health-related behaviours may be measured by direct observation.

Changes in knowledge or attitudes are sometimes an important initial step before a behaviour is changed, which, once changed, should reduce the risk of the disease of interest. Knowledge or attitudes can be assessed with reasonable reliability, using questionnaires or other interview methods, but observational studies may be required to determine if behavioural changes have actually occurred. For example, in a study to investigate the effectiveness of a health education campaign to promote the use of latrines, it may be relatively straightforward to assess, after the campaign, whether individuals have a better knowledge of why using latrines is desirable, but observational studies, before and after the campaign, may be necessary to ascertain whether or not the frequency of use of latrines had actually changed, let alone whether behavioural change led to a reduction in the incidence of diarrhoea. Similar issues arise with respect to the evaluation of a hand-washing intervention campaign. Further studies may then be needed to determine whether the changed behaviour has led to a reduction of diarrhoeal diseases.

Some trials have the incidence of a self-reported behaviour as one of their outcomes. For example, in evaluating the effectiveness of sexual behaviour change interventions, it is not possible to observe sexual behaviours directly, so self-reported behaviours are frequently recorded. But such measures are very open to desirability bias where the respondent reports the behaviour that they think the investigator would judge to be the desirable one. Furthermore, the desirability bias may be differential between the trial arms. For example, if the intervention group has been encouraged to reduce their number of sexual partners and always use a condom, while the control group has not, the intervention group may be more likely to over-report these 'desired' behaviours at follow-up. Self-reported behaviours, though sometimes the only practical outcome for a trial, are potentially misleading and should be avoided, at least as the primary outcome measure in a trial, if at all possible.

2.5.2. Transmission reduction

The purpose of interventions, based on vector control or environmental alteration, may be to reduce or interrupt transmission of the infectious agent of interest. Generally, the first priority is to determine whether the intervention has accomplished the immediate changes intended. For example, in trials in which insecticides are applied to reduce vector populations in order to reduce the transmission of some infectious agent, the first step would be to determine the impact of the intervention on the vector population. If the vector population is little affected, it may be reasonable to conclude that any impact on human disease is unlikely. However, if there is a reduction in vector population, it may be erroneous to conclude that the human disease load will also fall. A further study to determine the impact on disease may be required. Similarly, if interventions are being evaluated that may reduce indoor air pollution as a measure against respiratory disease, it may be best to focus initial studies on the assessment of changes in pollution levels, before assessing the impact on respiratory diseases. Usually, it will be more efficient to carry out trials to monitor the impact on disease only after there is evidence of an effect on the vector or on the agent against which the intervention is directed.

In order to assess a change in transmission, any, or all, of several different outcomes may be used:

- incidence of infection or disease

- prevalence of infection or disease
- severity of disease
- intensity of infection (for example, for helminths)
- intensity of infective agent in the vector.

Any changes to these different outcomes will happen at different intervals after the intervention is in place, and may require studies over time to measure the overall study impact. For instance, in an onchocerciasis control programme, the first evidence that an intensive larviciding of *Simulium damnosum* (black fly) breeding sites is having an effect may be a dramatic drop in fly-biting rates in the intervention area. Over the next several years, there may be a steady fall in the intensity of microfilarial infections among those living in the endemic area, but only after some years might it be possible to detect evidence of a fall in the prevalence of infection, and later still an impact on blindness rates which is the major adverse health consequence of onchocercal infection.

2.6. Adverse events

An important outcome of all trials is to assess the safety of the intervention under evaluation (for example, of a new drug or vaccine). Adverse events (AEs) are defined as any untoward clinical or laboratorial medical occurrence in a patient or clinical investigation subject, related or not to the use of an intervention in a trial. Serious AEs (SAEs) are defined as any events that are life-threatening or result in death. They include patient hospitalization or prolongation of existing hospitalization, events that result in persistent or significant debilitation or incapacity, and congenital anomalies and birth defects. All SAEs should be reported immediately to the sponsor (or the DSMB on behalf of the sponsor), followed by detailed written reports (see Chapter 7). Usually, two types of study outcomes are defined: (1) the active, prospective evaluation of a set of predefined potential AEs known or suspected to be associated with the type of drug, vaccine, or product under evaluation, and (2) recording all clinical or laboratory abnormalities, expected or not, that occur in study subjects during a specified time period or throughout the conduct of the trial, by active or passive surveillance, which may reveal an adverse consequence previously not known to occur with the drug, vaccine, or product under evaluation. For both types of safety outcomes, criteria must be developed to assess the severity, as well as the incidence of AEs associated with the drug, vaccine, or product under evaluation. Severity can be measured by the magnitude of a laboratory or clinical test abnormality, or by the subjective perception on how much the AE altered the function or quality of life of the individual. For instance, a reaction at the site of injection of a vaccine could be graded as mild if only a colour change is noted with mild pain, without induration and without any restriction on the arm or leg movement; moderate if, in addition to colour change of the skin, induration is noted and there is some restriction of movement; and severe if the subject cries out or winces if the area is touched and the arm or leg cannot be moved without pain. In many studies, a diary card may be provided to the study subject or, in case of children, to the mother or caretaker to record these reactions during a 7- or 14-day period after the administration of a vaccine or during the drug therapy. To aid measuring an injection site reaction, a ruler may be provided to the subject. And to standardize the measurement of temperature, a digital thermometer may be provided as well. Study subjects or children's mothers or caretakers need to be appropriately trained in using these study cards and instruments. In addition to its severity, these reactions are usually classified as unrelated, unlikely to be related, or possibly related to the intervention under evaluation. The criteria used for this classification may include proximity of the event to the administration of the intervention (for instance, a rash developing within 20 minutes of an injection would most likely be classified as possibly related), the unusualness of the clinical event (a disease which normally occurs in that age group or a complication expected to happen in the disease under study), or even the subjective interpretation of the investigator. Whatever criteria are used should be stated. The incidences of AEs, graded by the severity and likelihood of being related to the interventional product, are later compared between the study group exposed to the intervention and the control group (using placebo or an active comparator) to assess statistically if AEs of different kinds were or were not associated with the drug, vaccine, or product.

All safety measurements need careful definition in the study protocol, study forms to record them, using standardized measurements and codes to register them, and active monitoring of their occurrence. Most trials require those AEs that are considered serious to be individually reported to the sponsor and to an ethics review board, to the regulatory agency overseeing the trial, and to an independent DSMB for their careful evaluation during the conduct of the trial, to allow the possibility for the trial to be stopped or modified before its completion if it is suspected that SAEs are associated with the drug, vaccine, or product under investigation.

3. Factors influencing choice of outcome measures

The choice of the outcome measures in a specific trial largely depends on the purpose of the trial and how relevant, feasible, and acceptable the measures will be in a particular study population. Furthermore, the choice may be constrained by economic, logistic, or ethical considerations.

3.1. Relevance

Interventions are generally designed to reduce disease and/or to promote health. The outcome measures chosen should reflect these objectives as fully as possible, but, when intermediate variables are used, rather than those of main interest, care must be taken to choose variables of direct relevance to the main outcome. This is not always straightforward. For example, it may be decided to assess the impact of a vaccine by measuring the proportion of individuals who develop antibodies to the vaccine. This may be reasonable if it is known that there is a high correlation between the development of antibodies and protection from clinical disease. For many diseases, however, this relationship has not been established, and it would not be warranted to base conclusions regarding protection against disease simply on antibody determinations.

A health education intervention may be designed to change behaviour to reduce disease risk, but, as discussed in Section 2.5.1, asking individuals if they have changed their behaviour may give a measure of impact that correlates poorly with true changes in the risk of disease. Are individuals responding truthfully? Are they doing what they say they do? Even if behaviour changes, is this associated with a lowering in the incidence of disease?

The outcome variable measured should be as close as possible to the outcome of main interest. While this may seem an obvious suggestion, it may have major impact on the design of a study. For example, if the prevention of death is of prime interest, then, whenever possible, this should be made the endpoint of the trial. To do so might require an increase in the size of the trial from hundreds to thousands, or even tens of thousands, of individuals. Such a large trial might be difficult to find funding for, and there may never be an adequate test of whether the intermediate variables measured are acceptable surrogates for effects on mortality.

3.2. Feasibility

To be successful, a trial must be designed to have achievable objectives. A trial which has mortality as the endpoint, but which is too large to be successfully completed, may be of less value than a well-designed smaller trial aimed at assessing the impact on some intermediate endpoint such as severe disease. There must often be a compromise between relevance and feasibility. It is pointless to set unachievable goals, even if they look attractive in the objectives section of a proposal. Also, it may be of little value to measure the effect of an intervention on an outcome measure which is only distantly related to the measure of prime interest. The outcome measures selected will be much influenced by the resources available for the trial, the availability of skilled personnel, and the necessary laboratory support to diagnose cases of disease. In many large trials, every individual in the study population may have to be screened for disease or infection in a relatively short time. With such time constraints, some individuals may be misdiagnosed. The consequences of reductions in diagnostic sensitivity and specificity are discussed in Section 4.2.

3.3. Acceptability

The acceptability of the measurement of an outcome variable to the study population is critical to the successful conduct of a trial. For example, the recording of birthweights may not be possible in a population that allows only close relatives to have access to a mother for a few days or weeks after the child's birth. Taking venous blood samples or repeated blood samples is unpopular in many societies. If the method for measuring the outcome involves pain or inconvenience to the participants, it may be necessary to modify or abandon it. An outcome, of which the assessment involves a long interview with participants at a time when they would otherwise be planting crops or taking care of their household chores, may be unacceptable; it may either have to be abbreviated or carried out at a more convenient time.

3.4. Opportunity for add-on studies

Some trials offer the opportunity to measure outcomes that are not directly related to the objectives of the original study itself. These opportunities can be exploited by researchers to answer questions with minimal additional funding. For example, a diarrhoeal surveillance study might be carried out within a clinical trial in which a cohort of healthy children is being followed over time. However, it is very important that the add-on study does not interfere with the

original study outcome measure. Such additions should be considered at the beginning of the study and should have a separate study protocol. It is also important to inform sponsors, participants, and all stakeholders of the original trial of the coexistence of the proposed add-on study. Such investigations will usually require separate ethical approval and informed consent.

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 × 0.9 = 9.0 and 5 × 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 × 0.01 = 9.9) false cases = 19.9 per thousand in the unvaccinated group, and five true cases + (995 × 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.

References

Cutts, F. T., Zaman, S. M., Enwere, G., et al. 2005. Efficacy of nine-valent pneumococcal conjugate vaccine against pneumonia and invasive pneumococcal disease in The Gambia: randomised, double-blind, placebo-controlled trial. *Lancet*, **365**, 1139–46.10.1016/S0140-6736(05)71876-6 [PubMed: 15794968] [CrossRef]

Lopez, A. D., Lozano, R., Murray, C. J. L., Shibuya, K. 2011. Verbal autopsy: innovations, applications, opportunities improving cause of death measurement. *Population Health Metrics*, **9**, 128–254.

Ross, D. A., Changalucha, J., Obasi, A. I., et al. 2007. Biological and behavioural impact of an adolescent sexual health intervention in Tanzania: a community-randomized trial. *AIDS*, **21**, 1943–55.10.1097/QAD.0b013e3282ed3cf5 [PubMed: 17721102] [CrossRef]

Schellenberg, D., Menendez, C., Kahigwa, E., et al. 2001. Intermittent treatment for malaria and anaemia control at time of routine vaccinations in Tanzanian infants: a randomised, placebo-controlled trial. *Lancet*, **357**, 1471–7.10.1016/S0140-6736(00)04643-2 [PubMed: 11377597] [CrossRef]

World Health Organization. 2010. *International statistical classification of diseases and related health problems*, 10th revision, Volume 2, instruction manual [Online]. Geneva: World Health Organization. Available at: <http://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf>.

World Health Organization. 2012. *Verbal autopsy standards: the 2012 WHO verbal autopsy instrument* [Online]. Geneva: World Health Organization. Available at: <http://www.who.int/healthinfo/statistics/WHO_VA_2012_RC1_Instrument.pdf>.

Tables

Table 12.1 Examples of primary and secondary outcomes for trials of different interventions

Intervention trial	Primary outcome(s)	Secondary outcomes	Comment
Phase III trial of 9-valent conjugate pneumococcal vaccine in The Gambia (Cutts et al., 2005)	◆ First episode of radiological pneumonia	◆ Clinical or severe clinical pneumonia ◆ Invasive pneumococcal disease ◆ Invasive pneumococcal disease due to serotypes in vaccine ◆ All-cause hospital admissions ◆ All-cause mortality	The main purpose of the trial was to evaluate the public health impact of the vaccine. First episodes of radiological pneumonia were reduced by 37% (and all-cause mortality by 16%—not a primary endpoint in the trial). Highest efficacy was expected against invasive pneumococcal disease due to serotypes in the vaccine, but the aetiology of most cases of pneumonia is difficult to establish.

Intervention trial	Primary outcome(s)	Secondary outcomes	Comment
Cluster randomized trial to assess the impact of an adolescent sexual health intervention in Tanzania (Ross et al., 2007)	◆ Incidence of HIV infection ◆ Prevalence of herpes simplex type 2 (HSV 2) infection at end of trial	◆ Six biological measures (fo example, syphilis and gonorrhoea prevalence at end of trial) ◆ Five behavioural endpoints (for example, use of condoms during sexual intercourse) ◆ One attitudinal endpoint ◆ Three knowledge endpoints (for example, how HIV is transmitted)	The intervention was designed to reduce HIV incidence through behaviour change brought about by sexual health education. A substantial number of secondary outcomes were included to facilitate understanding of the main results. This was important, as the intervention was shown to substantially improve knowledge, reported attitudes, and some reported sexual behaviours but had no consistent impact on biological outcomes.
Trial of intermittent treatment of infants for malaria and anaemia control at time of routine vaccinations in Tanzania (Schellenberg et al., 2001)	◆ First or only episode of clinical malaria	◆ Multiple malaria episodes ◆ Fever episodes ◆ Severe anaemia ◆ Admissions to hospital ◆ Outpatient attendances	This was a test of a new approach to malaria control by administering anti-malarial drugs routinely to infants attending clinics for vaccination. Clinical malaria was reduced by 59%, and severe anaemia by 50%

Bookshelf ID: NBK305519