12.2: 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.