2.12 Medical screening 137

2.12 Medical screening 137

ESSENTIALS Medical screening is the systematic application of a test or inquiry to identify individuals at sufficient risk of a specific disorder to benefit from further investigation or direct preventive action (these individ- uals not having sought medical attention on account of symptoms of that disorder). Key to this definition is that the early detection of disease is not an end in itself; bringing forward a diagnosis without altering the prognosis is useless and may be harmful. Criteria for screening Before a potential screening test is introduced into practice it must be shown to prevent death or serious disability from the disease to an ex- tent sufficient to justify the human and financial costs. To this end, three screening parameters need to be determined: (1) the detection rate (sensitivity); (2) the false-positive rate (equivalent to the specificity); and (3) the odds of being affected given a positive screening result (equiva- lent to the positive predictive value). Where a detection rate cannot be directly determined (e.g. in cancer screening, or if the efficacy of the intervention is uncertain), a randomized trial is needed to show that screening and subsequent treatment reduce disease-specific mortality. Circumstances where screening is not appropriate Screening tests should not be practised simply because they seem intuitively useful: chest radiography to screen for lung cancer and manual breast self-examination to screen for breast cancer were as- sumed to be worthwhile, but randomized trials showed they did not significantly reduce mortality from the cancer. Screening for prostate cancer is widely practised, yet it does harm (from hazardous treat- ment) with evidence of a relatively modest reduction in mortality from the disease. Causal risk factors, even important ones like serum cholesterol and blood pressure for cardiovascular disease, usually dis- criminate poorly between individuals who will and will not develop the disease they cause, because most of the population is ‘exposed’. Particular disorders where screening is justified The number of disorders for which medical screening has been shown to be worthwhile is perhaps surprisingly small, but includes: (1) antenatal screening (e.g. various single-gene disorders, Down’s syndrome, neural tube defects, and some infections such as hepa- titis B and HIV that may be asymptomatic in the mother but cause disease when transmitted to the fetus); (2) neonatal screening (e.g. congenital hypothyroidism, certain inborn errors of me- tabolism such as phenylketonuria, and congenital deafness); (3) adult screening—in respect of cancer this has been shown to be worthwhile for only three cancers—breast, cervical, and colorectal; screening individuals with diabetes mellitus prevents blindness from retinopathy; screening men around the age of 65 prevents death from ruptured abdominal aortic aneurysm; and screening young women for chlamydia infection prevents pelvic inflamma- tory disease and its complications (including infertility). Future prospects Tests that arise out of technological development in the absence of a clear case of medical need (e.g. whole-body scanning using MRI or CT), should not be ‘sold’ to the public in the belief that they are

helpful. As with all screening methods, their value needs to be shown before they are introduced into practice. Determining when medical screening is an effective method of preventing serious dis- ease and disability is one of the most challenging areas in medical research. Introduction There is scarcely a medical discipline that does not include some aspect of screening. It has made significant inroads into the pre- vention of disease, but is often used inappropriately in circum- stances where there is insufficient evidence that it benefits health. Determining when screening is an effective method of prevention is one of the most challenging areas in medical research today, re- quiring an understanding of the principles of screening, the path- ology, natural history, and epidemiology of the diseases concerned, and quantitative information on the efficacy of the screening tests and the remedies available. 2.12 Medical screening Nicholas Wald and Malcolm Law1 1 © The author.

138 section 2 Background to medicine Medical screening contains three elements: 1. Identifying individuals at sufficiently high risk of having or developing a specific disorder to benefit from further investiga- tion or direct preventive action. 2. It is systematically offered to a population that has not sought medical attention for symptoms of the relevant disease. It is usu- ally initiated by medical authorities, not the patient. 3. Its purpose is to benefit screened individuals. On this basis, mass testing activities such as surveillance for HIV infection or pre- employment examinations to test fitness for work are not classi- fied as medical screening. The following definition has been widely used and encapsulates these three elements: Medical screening is the systematic application of a test or inquiry, to identify individuals at sufficient risk of a specific disorder to benefit from further investigation or direct preventive action, among those who have not sought medical attention on account of symptoms of that disorder. Worthwhile screening aims to prevent death or disability from specific disorders. Screening that simply brings forward a diag- nosis without altering the prognosis is useless and may be harmful, prompting needless anxiety, and possibly hazardous interven- tions. The early detection of disease is not an end in itself. As with any medical treatment, screening needs to be shown to offer medical benefit and to be acceptably safe. Many medical disorders are not candidates for screening, because they are too trivial or because treatment is no more effective following screening than following clinical presentation. The value of a screening test, in which the benefits are considered in the light of the human and financial costs, needs to be determined before it is introduced into practice. Requirements for a worthwhile screening test See Box 2.12.1. The disorder The disorder needs to be clinically well defined and should, wher- ever possible, be specified independently of the screening test. The disorder should not be an ‘abnormal’ value of the screening test being offered, such as a value lying outside the 95% range. This cre- ates a circularity (‘tautological screening’) that makes it impossible to determine whether screening is genuinely preventing disease or is just causing overdiagnosis. It is necessary to know the distribution of values of the screening test in individuals who have (or will develop) the clinical disease and in individuals who do not, in order to assess the value of the test. An example is hypertension, or high blood pressure, an asymp- tomatic condition that increases risk of a heart attack or stroke. If hypertension were regarded as the medical disorder being screened for, then all the ‘hypertensives’ would have blood pressure above the cut- off and all the ‘nonhypertensives’ below it (i.e. a perfect test). The apparent screening perfection is a tautological misconception. A high blood pressure measurement is the result of a screening test (blood pressure measurement) for the clinical diseases (stroke and myocardial infarction) caused by high blood pressure. In fact, blood pressure measurement, although widely practised, is not a good screening test for stroke or myocardial infarction. Many people who will not have a stroke or

myocardial infarction have high blood pressure, and many who do will not. This is considered in further detail next. Prevalence or incidence To derive an estimate of the odds of being affected among indi- viduals with a positive screening result (see next), the prevalence, or incidence, of the disorder needs to be known. Prevalence is the number of cases of a disorder in a defined population at a given point in time, incidence is the number of new cases occurring in a de- fined population over a specified period. If the disorder is very rare screening may not be justifiable, unless it can easily be incorporated into an existing screening protocol. If the disorder is very common (e.g. heart attacks and strokes), screening may be pointless and a population-wide preventive strategy may be needed. Natural history Screening should be restricted to disorders that are medically important (i.e. associated with serious morbidity or premature mortality). Remedy A remedy or treatment must be available that is more effective or ac- ceptable following screening than at clinical presentation. Offering an effective treatment is insufficient; the treatment must be more effective or acceptable if delivered early. Screening test The screening test should be simple and safe. Some screening tests are so simple or performed so routinely that they are not recognized as such. For example, asking a woman's age was once the antenatal screening test for Down's syndrome. A routine blood count includes Box 2.12.1 Requirements for a worthwhile screening test 1 Disorder: well defined 2 Prevalence/incidence: known 3 Natural history: medically important disorder 4 Remedy or treatment: more effective or acceptable than at clinical presentation 5 Screening test: simple and safe 6 Test performance specified: (a) Detection rate can be determined: detection rate and false- positive rate known and acceptable. For a quantitative screening test, the distributions of test values in affected and unaffected individuals should be known, the extent of overlap sufficiently small, and a suitable cut-off level defined (b) Detection rate cannot be determined: randomized trial evidence shows that the combined effect of screening and treatment is sufficiently effective in preventing death and disability from the disease being screened for, with an acceptably low proportion of individuals requiring further investigation 7 Financial: overall cost acceptable to achieve the health benefit 8 Facilities: available or can easily be installed, including for diagnosis and treatment 9 Acceptability: procedures following a positive result are generally agreed and acceptable to the screening authorities and the screened individuals

2.12 Medical screening 139 the antenatal screening test for β-thalassaemia (mean corpuscular volume), so the

issue is not one of introducing the test but of system- atically interpreting a test already carried out. The purpose of testing generally defines whether it is a screening or diagnostic test. If the aim is to identify a high-risk group for fur- ther investigation or preventive treatment, it is a screening test; if it is to make a diagnosis, it is a diagnostic

test. Screening tests indicate a probability of having or developing a disorder, whereas diagnostic tests usually indicate whether an individual is affected or unaffected. The accuracy of each type of test does not itself define what type of test it is. Sometimes mass testing, perceived as screening, is in fact diagnosis (e.g. obstetric

ultrasonography used routinely to diagnose anencephaly). Screening tests usually apply to healthy populations, but this is not always the case (e.g. screening for retinopathy among people with diabetes). Screening test performance It is useful to separate screening tests for which detection rates can be determined from

screening tests for which this is not possible. Detection rate can be determined The performance of screening and diagnostic tests is defined by three parameters: (1) the detection rate; (2) the false-positive rate; and (3) the odds of being affected given a positive result (OAPR). Detection rate The detection

rate of a test (or test sensitivity) is the proportion of affected individuals with positive test results (Table 2.12.1). An advantage of the term detection rate over sensitivity is that it avoids confusion with the usage of sensitivity in analytical biochem- istry, where it means the minimum detectable amount in an assay. In cancer screening,

‘detection rate’ is often taken to mean the prevalence of detected cancers at a screening examination. False-positive rate The false-positive rate is the proportion of unaffected individuals with positive test results (Table 2.12.1). The complement of the false-positive rate is the specificity, which is 100% minus the false-positive rate

(e.g. a false-positive rate of 3% is the same as a specificity of 97%). Advantages of the term false- positive rate over specificity are that (1) it is more easily understood and remembered; (2) it focuses attention on the group to be offered further medical intervention; and (3) a 10% false-positive rate is twice as ‘bad’ as one of 5%, whereas

this is concealed within the cor- responding specificity values of 90% and 95%. Odds of being affected given a positive result (OAPR) The OAPR is the ratio of the number of affected to unaffected indi- viduals among those with positive test results (i.e. true positives:false positives in the population in question). The OAPR in Table 2.12.1

would be a:b if the numbers in Table 2.12.1 came directly from screening everyone in a study popu- lation. In practice this is uncommon because the disorder being screened for is rare and so it is sensible to estimate the detection rate on all the affected individuals but only a small sample of un-affected individuals. Because of this sampling

difference, tables like Table 2.12.1 cannot be used to estimate the OAPR. It is best esti- mated indirectly using estimates of the prevalence of the disorder from one source and estimates of the detection rates and false-positive rates of the screening test from another source. This can be done using a flow diagram such as that in Fig. 2.12.1, in which

the detection rate (80%) is applied to the number of affected individ- uals (prevalence 1%) and the false-positive rate (4%) to the number of unaffected. Then the ratio of true-positive to false-positive tests performed will be an unbiased estimate of the OAPR in a total population. The OAPR is 1:5 after the screening test, and 38:1

after the diagnostic test (detection rate 95%, false-positive rate 0.5%). If the prevalence of the disorder were halved (0.5%) the OAPRs would be halved to 1:10 and 19:1, respectively. Thus, the less common the disorder, the less likely people with positive results will be affected. The OAPR can be expressed as a probability (‘true

positives/all positives’) which is known as the positive predictive value (PPV). In the example in Fig. 2.12.1) the OAPR, 80:400 = 1:5, is equivalent to a predictive value of 1/(1+5) = 1/6 - 17%. The OAPR is more useful than the PPV because it is numerically easier to compute when tests are performed in sequence (Fig. 2.12.1), and it provides a

better impression of the relative performance of tests. In the example, the OAPR of 38:1 is equivalent to a PPV of 97% (38/39). If the detection Table 2.12.1 Algebraic summary of detection and false-positive rates of qualitative tests or quantitative tests using a specified cut-off Test result Affected Unaffected Positive TRUE POSITIVES a

FALSE POSITIVES b Negative FALSE NEGATIVES c TRUE NEGATIVES d Total a + c b + d Detection rate (sensitivity) a a+ c False-positive rate (1 – specificity) b b+ d 10000 individuals 100 affected Screening test Diagnostic test DR = 80% DR = 95% 80+ve 400+ve OAPR = 80:400 1:5 76:2 38:1

9900 unaffected 76+ve 2+ve FPR = 0.5% FPR = 4% Fig. 2.12.1 Flowchart to show the performance of screening and diagnostic tests. The critical first step in constructing such a flowchart is to separate individuals into affected and unaffected, not into

screen-positive and screen-negative. Reproduced from Wald, N. An Introduction to Epidemiology in Medicine. London: Royal Society of Medicine Press, 2004.

140 section 2 Background to medicine rate of the screening test were halved (to 40%) the OAPR would also be halved (to 19:1) but the PPV, 95% (19/20), appears only a little lower. A good screening test has a high detection rate, a low false-positive rate, and a high OAPR (e.g. 1:10 is better than 1:50). Stating the de- tection rate for a test is uninformative unless a false-positive rate (or specificity) is also stated. Screening performance is assessed by spe- cifying the detection rate for a given false-positive rate, or specifying the false-positive rate for a given detection rate. The detection rates and false-positive rates are independent of the prevalence of the disease for tests that measure a consequence of the disease (e.g. the antenatal markers of Down's syndrome) but may not be independent for a screening test that is a measure of a cause of the relevant disease. For example, when screening for an autosomal recessive disease such as cystic fibrosis by testing for a known DNA mutation in the gene for the disease, a higher gene prevalence will necessarily be linked to a higher disease prevalence. The OAPR is always dependent on the prevalence. The higher the prevalence the higher the OAPR, even if the detection rate and false- positive rate are constant. Estimates of the detection rate and false-positive rate can be ap- plied from one population to others because they are generally inde- pendent of the prevalence of the disorder. This is not the case with the OAPR, which depends on the prevalence. For a qualitative (or categorical) test, such as the presence or ab- sence of a cystic fibrosis mutation among a given panel of muta- tions tested for, there is only one detection rate and false-positive rate. This is not the case with quantitative (or noncategorical) tests, such as maternal serum α-fetoprotein (AFP) for open spina bifida screening, which yield numerical results. In such cases, the detection rate and false-positive rate depend on the screening cut-off level used to distinguish positive from negative results. No single pair of detec- tion and false-positive rates summarizes the performance of tests; both will vary as the cut-off is changed. For example, at cut-off level A in the relative frequency distributions in Fig. 2.12.2 the test will have a detection rate given by the area under the curve for affected subjects to the right of cut-off level A (95%), and a false-positive rate given by the area under the curve for unaffected subjects to the right of the same cut-off level (10%). The higher the cut-off level (say, B or C) the lower the detection rate and false-positive rate. It is common to summarize the performance of a test as a receiver– operator characteristic (ROC) curve. This is a plot of the detection rate against the false-positive rate, with both scales plotted from 0% to 100%. In such a graph a useless test is represented by the diagonal, indicating that the detection rate and the false-positive rate are al- ways the same. As the screening test improves, the ROC curve bows out from the diagonal towards the axes. A perfect screening test clings to the detection rate axis up to 100% while the false-positive rate remains at zero. The area under a ROC curve is sometimes used to indicate the performance of a screening test, but it is not a satis- factory measure of this. It is better to state detection rate for specified false-positive rate or vice versa. A weakness of a ROC curve is that for screening tests that are potentially useful, the area of the graph that is informative is restricted to a small portion, namely the part covering false-positive rates up to about 10% and detection rates (from 40% to 100%) between about 50% and 100%. Fig. 2.12.3 illus- trates detection rates plotted against false-positive rates (from 0% to 10%) in multiple marker antenatal screening for Down's syndrome, showing the improvements in screening that have been made over the past 20 years. Good screening tests are usually early manifestations of the dis- ease being screened for, while causes of a disease that are highly prevalent in a community are usually poor screening tests. Causal risk factors such as blood pressure for stroke are important aetio- logically

and account for a large proportion of the disease they cause because they are usually common (e.g. most adults over 55 can be said to have a high blood pressure) yet many escape the consequences (e.g. a stroke). This means that causal risk factors usually do not discriminate well between individuals who will and who will not develop the disease. Table 2.12.2 shows the detection rate for a 5% false-positive rate (DR5) for various risk ratio estimates between the top and bottom fifths of the distribution of a risk factor. Even a ‘strong’ risk factor with a fivefold risk ratio between the top and bottom quintile groups (fifths) of the distribution (typical of low-density lipoprotein (LDL) cholesterol and myocardial infarction) has only a 14% detection rate Affected A Unaffected Cut-off A at 6 units DR FPR 95% 10% Cut-off B at 6.5 units DR FPR 90% 2.5% Cut-off C at 7 units DR FPR 80% 1% Unaffected Unaffected Test variable (arbitrary units) 2 3 4 5 6 7 8 9 10 11 B C Affected Affected Fig. 2.12.2 Hypothetical example of the detection rate and false- positive rate of a screening test at three different cut-off levels. The implied vertical axis is the percentage of individuals at different levels of the screening test variables, considered separately for affected and unaffected individuals. Reproduced from Wald, N. An Introduction to Epidemiology in Medicine. London: Royal Society of Medicine Press, 2004.

for a 5% false-positive rate. An interquintile risk ratio of around 1000 is necessary to achieve a detection rate of at least 75% for a 5% false-positive rate. The OAPR can be determined by using the

flowchart method il- lustrated in Fig. 2.12.1. It can also be determined using the likeli- hood ratio (LR) which is a measure of the ‘concentrating’ power of a test (Fig. 2.12.4). For a group of people with values of the screening variable above a specified cut-off (i.e. all screen positives), this is the proportion of the area for ‘affecteds’ to the right of

the cut-off div- ided by the proportion of the area for ‘unaffecteds’ to the right of the cut-off. Fig. 2.12.4a, which is equivalent to the detection rate divided by the false-positive rate (DR/FPR). It is the number of times indi- viduals with positive results are more likely to have the disorder for which they are being tested compared with the general

population (individuals who have not been tested). That is, the OAPR is the likelihood ratio multiplied by the preva- lence of the disorder (expressed as an odds): OAPR LR prevalenceasanodds.

× So (see Fig. 2.12.4a for example), if the detection rate is 80% and the false-positive rate is 1%, then the LR is 80%/1%, or 80. If the

prevalence of the disorder were 1:1000, then OAPR 80 1:1000 80 :1000 1:1000 / 80 1:12.5.

×

=

For an individual with the screening variable at some specific value, the likelihood ratio is the height of the relative distribution curve for ‘affecteds’ at the test value for that individual divided by the height of the curve for ‘unaffecteds’ at the same test value. So, for example, an 100 90 80 60 50 40 0 1 2 3 4 5 False-positive rate (%) Integrated test 9.4 8.6 Serum integrated test 8.7 7.3 Combined test 8.6 7.2 Quadruple test 8.3 6.4 Triple test 6 7 8 9 10 70 Detection rate (%) 7.7 5.6 Fig. 2.12.3 Antenatal screening for Down’s syndrome: detection rates and false-positive rates for specified screening tests. The integrated test consists of the ultrasound marker nuchal translucency and pregnancy- associated plasma protein A (PAPP-A) measured in the first trimester, and AFP, unconjugated oestriol (uE3), human chorionic gonadotropin (hCG), and inhibin-A in the second trimester. The serum integrated test is the same as the integrated test without nuchal translucency. The combined test consists of nuchal translucency, PAPP-A, and hCG in the first trimester. The quadruple test consists of AFP, uE3, hCG, and inhibin-A in the second trimester. The triple test is the same as the Quadruple test without inhibin-A. All the tests include maternal age. Reproduced from Wald NJ, et al. (2004). SURUSS in perspective. Br J Obstet Gynaecol, 111, 521–31, with permission from Wiley-Blackwell. Table 2.12.2 Detection rate for a 5% false-positive rate (DR5) according to relative risk between top and bottom fifths of the distribution in unaffected individuals Relative risk between top and bottom fifths of
the distribution in unaffected individuals DR5 (%) 1 5 2 8 3 11 5 14 10 20 40 36 80 45 800 71 2000 79 10 000 89 Reproduced from Wald NJ, Hackshaw AK, Frost CD, ‘When can a risk factor be used as a worthwhile screening test?’, BMJ, 319, 1562–65 © 1999 with permission from BMJ Publishing Group. Affected Affected a b Unaffected Unaffected (a) Likelihood ratio for groups (b) Likelihood ratio for individuals DR = 80% FPR = 1% LR = 80%/1% = 80 FPR = 1% LR = a/b = 12/1

= 12 DR = 80% Test variable (arbitrary units) 2 3 4 5 6 7 8 9 10 11 12 Fig. 2.12.4 Likelihood ratio for groups and for individuals. Reproduced from Wald, N. An Introduction to Epidemiology in Medicine. London: Royal Society of Medicine Press, 2004.

142 section 2 Background to medicine individual with a test result of 7 (arbitrary units) in Fig. 2.12.4b has a likelihood ratio of 12, and so OAPR 12 1:1000 12 :1000 1:1000 /12 1: 83.

×

=

In this way the likelihood ratio is used to estimate the risk for an individual. Fig. 2.12.5, showing the distribution of diastolic blood pressure in men who did and did not subsequently die of a stroke, illustrates how a particular blood pressure measurement of, say, 105 mm Hg in a 70-year-old man, can be converted into a risk of developing a stroke. At a diastolic blood pressure of 105 mm Hg the likelihood ratio is 3. The annual risk of a fatal stroke in all 70-year-old men regardless of blood pressure is 2:1000 (about 0.2%), so if his dia- stolic blood pressure is 105 mm Hg the risk is 3 × 2:1000 or 6:1000 (about 0.6%). To establish whether a quantitative screening test is worthwhile,

the overlapping distributions of the values of the screening test in people with and without the disorder must be examined. If the two distributions are widely separated, as in the example in Fig. 2.12.6 (ultrasound measurement of the diameter of the abdominal aorta as a screening test for aneurysm likely to rupture), the test is good. If they substantially overlap, as in the example in Fig. 2.12.7 (serum cholesterol as a screening test for future death from ischaemic heart disease or blood pressure as a screening test for stroke, Fig. 2.12.5), it is not. Detection rate cannot be determined Determining the detection rate is straightforward when all indi- viduals can be found to be either affected or unaffected. This is not always possible, notably in cancer screening, because if a lesion is found and a treatment carried out, one cannot know if that lesion would have become a clinical case had treatment not been given, or if it is 'overdiagnosis'. The problem arises for any progressive dis- order in which the clinical outcome is not determined in a uniform way among all individuals, as would be the case if all screening re- search were initially observational without intervention dependent on the result of the screening test. Sometimes such an observational approach is possible, for example, by storing serum samples in a population of adults (without testing them at the time of collection), and later identifying those who did and did not develop a cancer, retrieving the serum samples, and testing them on a case–control basis; this provides an unbiased estimate of the screening perform- ance of the test. Such an approach is not practical with tests based on imaging, such as mammograms, which could not ethically be taken and stored without being examined at the time. In such circum- stances, it may never be possible to know the screening performance of the test. The solution is to perform a randomized trial of screening (and treatment) versus no screening. If this shows that mortality from the disease is reduced, the combined effect of screening and treatment is known, though the relative contributions of the two in achieving the health benefit may not be. Cancer screening must prolong survival (the time between diag- nosis and death) to be effective, but because of two biases, prolonged survival alone is insufficient evidence that screening genuinely Men who die of stroke 3 1 Men who do not die of stroke Diastolic blood pressure (mm Hg) 40 50 60 70 80 90 100 110 120 130 140 150 Fig. 2.12.5 Likelihood ratio of a fatal stroke in a man with a diastolic blood pressure of 105 mm Hg. Reproduced from Wald, N. An Introduction to Epidemiology in Medicine. London: Royal Society of Medicine Press, 2004. Ruptured aneurysms (n = 163) Unruptured aortas (n = 3897) Maximum aortic diameter (cm) 0 2 4 6 8 10 12 14 16 Fig. 2.12.6 Aortic diameter and ruptured aortic aneurysm. The distribution of less than 2 cm is not real but simply represents the lower half of the Gaussian distribution of more than 2 cm, which is based on data. Data from Law MR, Morris J, Wald NJ (1994). Screening for abdominal aortic aneurysms. J Med Screen, 1, 110–116. Died of IHD Did not die of IHD Serum cholesterol (mmol/litre) 15% 5% 2 3 4 5 6 7 8 9 10 11 Fig. 2.12.7 Relative distributions of serum cholesterol in men who subsequently died of ischaemic heart disease (IHD) and in men who did not. Reproduced from 'A strategy to reduce cardiovascular disease by more than 80%', NJ Wald & MR Law, British Medical Journal 2003, 326; 7404 with permission from BMJ Publishing Group Ltd.

 improves prognosis. The first bias, lead time bias, is the prolongation of survival from bringing forward the date of diagnosis, even though the date of death is unchanged. The second bias, length time bias, arises because cancer screening involves periodic examinations (say 3-yearly). So screening will detect slowly growing tumours more readily than rapidly growing ones because rapidly growing ones are more likely to develop and proceed to clinical presentation within the interval between two consecutive screening examinations, and thereby escape detection at screening. Survival with such rapidly growing screen-detected cancers will inevitably be shorter

than average. This is biased sampling. Both biases can be avoided by com- paring mortality from the specific cancer (the number of deaths divided by the number of people at risk) between screened and unscreened groups rather than comparing survival once a cancer is diagnosed. The biases are avoided because mortality measures deaths, whereas survival measures time. Disease-specific mortality could be subject to bias if the screened and unscreened groups were at different risks of developing the disease. For example, women of higher socioeconomic status may be more likely to develop breast cancer and more likely to ac- cept screening. So breast cancer mortality could still be higher in screened women even if screening were effective. The only way to reliably avoid such selection bias is to carry out a randomized con- trolled trial to be sure that like is compared with like. Financial considerations Having determined that the first six requirements for a worthwhile screening test are met (see Box 2.12.1), the financial considerations need to be assessed. Screening programmes should seek to minimize the cost for a given outcome (i.e. to maximize the cost-effectiveness). If the most medically effective form of screening is also the most cost- effective, it should be the programme of choice, provided it is affordable. If the best screening policy is not the most cost-effective, a judgement is needed on whether the extra health gain justifies the extra cost. Facilities Medical screening effectively 'creates' patients by identifying indi- viduals at sufficient risk of a disorder to be offered further tests or treatment when they had no prior suggestion that they may have the disorder. This necessarily creates anxiety and a demand for medical attention, and an obligation to ensure that facilities exist for the ne- cessary investigation, treatment, and support. Screening should not be implemented until such arrangements have been made. Screening therefore needs to be offered in the context of programmes that are capable of meeting all the related needs of the people being screened. Acceptability Medical screening, including the treatment or remedy, must be ac- ceptable to the population concerned and to the professional staff involved. The purpose, the benefits, and the limitations of screening need to be understood and regarded as important from the perspec- tive of each individual who is offered screening. The decision not to be screened needs to be respected and programmes should not be driven by targets that set high uptake rates, though of course, if the rates are very low it would call into question the need for the screening programme. A key element in the acceptability of screening is individual choice set against a justifiable trust in the medical system that offers screening. Requirements for a worthwhile public health screening programme See Box 2.12.2. Screening is a public health activity that should meet certain re- quirements that arise from a professional responsibility to achieve a collective health benefit. It is not the provision of a consumer commodity. Its purpose is to improve the health of individuals and thereby the health of the community. Once the requirements for a worthwhile screening test shown in Box 2.12.1 are met, there are four additional requirements for a worthwhile screening programme implemented as a public health service. These are summarized in Box 2.12.2. In public health terms, interventions that reduce exposure to the causes of disease should have priority over screening to detect early disease and offer treatment, but they are not mutually exclusive. A population approach correcting or reversing adverse risk factors is often more effective. For example, the human papillomavirus (HPV) vaccine is expected to steadily replace screening for cervical cancer after the next 40 years. Screening for specific disorders Antenatal screening See Table 2.12.3. Much of antenatal care is screening—looking for problems be- fore they arise clinically. Detecting rises in blood pressure to warn of the risk of pre-eclampsia (which may cause perinatal death and serious illness in the mother), and detecting maternal anti-D anti- bodies to warn of rhesus haemolytic disease of the newborn, are two examples. The purpose of such screening is usually the welfare of the mother and fetus, but in antenatal screening there is the

unusual situation in which some fetal disorders are so severe or potentially disabling to justify screening and diagnosis, and the offer of a ter- mination of pregnancy. Antenatal screening for open neural tube defects, Down's syndrome, severe congenital heart malformations, and severe, incurable single-gene disorders are examples. Screening for four infections is worthwhile (syphilis, HIV, hepa- titis B, and bacteriuria) because they may not be clinically apparent in the mother but can cause serious preventable illness in the neo- nate (either immediately or in later life). Prognosis is substantially improved if the infection can be detected in the mother and Box 2.12.2 Requirements for a screening programme implemented as a public health service 1 Equitable: equal access to screening services 2 Organized: individuals are offered screening in an organized manner according to a specified protocol and with relevant information pro- vided to permit an informed choice 3 Comprehensive: screening is the first step in a programme of service and care that includes counselling screen positives, diagnosis, support, and treatment 4 Monitored and auditable: key aspects of the programme should be monitored so that remedial steps can be taken if they are below standard

144 section 2 Background to medicine Table 2.12.3 Summary of antenatal screening tests of proven value Disorder Approximate natural birth prevalence (per 10 000) in UK Primary screening test Secondary screening test(s) (if available) Detection rate (%) False- positive rate (%) Odds of being affected given a positive result Diagnostic test Intervention 1°ry screening test 2°ry screening test Autosomal or sex-linked recessive disorders Cystic fibrosis 4 Test for CF mutation in both parents ('couple screening') 72 0.09 1:3 – CVS or amniocentesis a Sickle cell disease 3 Ethnic origin enquiry (Afro-Caribbean) Sickling test; Hb electrophoresis in mother, and in father if positive in mother 99 3 1:100 1:3 CVS or amniocentesis a β-Thalassaemia 6 Red cell MCV or MCH in mother Hb A2 assay in mother, and in father if positive in mother 89 7 1:125 1:3 CVS or amniocentesis a Tay–Sachs disease 0.04 Ethnic origin enquiry (Ashkenazi Jew) Hexoseaminidase assays in father, and mother if positive in father 50 1 1:3600 1:3 CVS or amniocentesis a Haemolytic disease of the newborn (D-antigen of Rh system) 40 Rh grouping and test for antibody in mother Rh grouping of father; quantitation of maternal antibody 100 16 1:31 1:26 CVS or amniocentesis Intrauterine transfusion, early delivery with exchange transfusion Haemophilia 0.5 Recognition of affected male relative (carrier detection) Test for mutation in mother 55 <0.01 1:35 1:3 CVS or amniocentesis a Chromosomal disorders Down's syndrome (Trisomy 21) 18 Integrated 1st and 2nd trimester 89 1.0 1:5 CVS or amniocentesis a 1st trimester alone 77 1.0 1:6 CVS or amniocentesis a 2nd trimester alone 70 1.0 1:6 CVS or amniocentesis a Reflex DNA screening 91 0.05 4:1 Amniocentesis Trisomy 18 2.3 Integrated 1st and 2nd trimester 92 0.2 1:3 CVS or amniocentesis 1st trimester alone 89 0.2 1:3 CVS or amniocentesis 2nd trimester alone 58 0.2 1:5 CVS or amniocentesis Reflex DNA screening 89 0.05 1:1 Amniocentesis Trisomy 13 1.4 Integrated 1st and 2nd trimester 72 0.2 1:12 CVS or amniocentesis 1st trimester alone 73 0.2 1:11 CVS or amniocentesis 2nd trimester alone 20 0.2 1:42 CVS or amniocentesis Reflex DNA screening 79 0.05 1:3 Amniocentesis

2.12 Medical screening 145 Other congenital malformations Spina bifida (open) 8.5 Maternal serum AFP assay Ultrasound 87 0.5 1:4 Amniotic fluid acetylcholinesterase + repeat ultrasound a Anencephaly 10 Ultrasound 100 0 1:0 – Independent conf irmation a Severe cardiac malformations 20 Ultrasound 46 ≤0.6 ≥1:6 – Independent conf irmation a Infections transmitted from mother to fetus Congenital rubella syndromeb 0.12 Absent antibodies in mother

90 1.6 <1:1300 – None Vaccinate mother after delivery to protect subsequent pregnancies Congenital syphilis 0.2 VDRL test or flocculation test in mother Specific treponemal test in mother 90 0.2 1:100 1:50 None Penicillin AIDS 1 ELISA test for IgG antibody in mother (repeated on same sample if positive) ELISA test on repeat sample 99.9 0.13 1:13 1:<5 None Antiretroviral drugs to mother and infant Hepatitis B causing hepatoma and chronic liver disease 1.4 ELISA test for HBsAg in mother (repeated if positive) ≥98 0.14 1:10 None Recombinant vaccine to neonate, hepatitis B immunoglobin at birth except when mother has antibodies to e antigen Maternal bacteruria causing pyelonephritisc 200 Urine culture 76 4 1:4 None Antibiotics to mother Noninfectious maternal disease affecting fetus Maternal high blood pressure/ pre-eclampsia causing perinatal death 93 (rate of all perinatal deaths) Maternal blood pressure measurement Test for proteinuria 38 (of all perinatal deaths) 30 1:77 1:41 None Blood pressure lowering drugs a Information on disorder and its prognosis, counselling, termination of pregnancy, or preparation for birth of affected child, advice on risk of recurrence. b Worthwhile only with low uptake of childhood rubella vaccination in community. c May cause low birthweight or fetal death. AFP, α-fetoprotein; CVS, chorionic villus sampling. Adapted from Wald NJ, Leck I, eds. Antenatal and Neonatal Screening (2nd ed). (2000) Oxford University Press, Oxford.

146 section 2 Background to medicine appropriate treatment given to the mother before birth, the neonate at birth, or both. Routine screening for rubella syndrome in preg- nancy is generally not worthwhile because it cannot prevent the dis- order in the pregnancy screened; it can only lead to vaccination after birth in women without rubella antibodies. The preferred method of prevention is childhood vaccination. In recent decades antenatal screening has taken on a scientific methodology and rigour that has permitted the development of screening programmes that are now standard throughout the world. The first such initiative arose with antenatal screening for open neural tube defects, first by measurement of maternal serum AFP and later by ultrasonography, which is used with AFP in many places and has replaced it in some. Screening now detects virtually all cases of anencephaly with scarcely any false positives, and 87% of cases of open spina bifida with a false-positive rate of less than 1%. The birth prevalence of neural tube defects in Britain has declined by over 90% from 1 in 250 births in the early 1970s to less than 1 in 2500 now, due in part to screening, in part to an increase in folate intake through food and vitamin supplements. Until the 1980s, antenatal screening for Down’s syndrome (trisomy 21) was based on maternal age. In 1988 the triple test was described, based on combining second trimester serum markers with maternal age. Fig. 2.12.3 shows the subsequent improvement in screening per- formance as the number of available markers increased over time. The integrated test can detect about 85% of affected pregnancies for a false-positive rate of only 0.9%; the low false-positive rate is important because women with positive results usually have an amniocentesis, which may induce the miscarriage of a healthy fetus. Combining markers to obtain a single test result for an individual involves the multiplying of the likelihood ratios for each marker in that individual (as in Fig. 2.12.4a), allowing for any correlation between them (con- sidered separately among affected and unaffected individuals). So, for example, in the simple situation of three independent screening

markers that correspond to likelihood ratios of 3, 4, and 5, the com- bined likelihood ratio is 60 (3 × 4 × 5). To determine the screening performance of tests based on multiple markers, a hypothetical popu- lation of screened individuals is generated and the combined like- lihood ratio for each individual calculated and converted to risk by multiplying it by prevalence expressed as an odds. The overlapping distributions of risk in affected and unaffected individuals are plotted, determining detection rates for specified false-positive rates in the same way as for a single screening marker. Then risk itself becomes the screening variable—which is convenient, because it is exactly what is needed in reporting results to screened individuals. Most screening markers associated with Down's syndrome vary with gestational age, so a high level at one gestational age could be low at another. A widely used advance in screening is to express all values as 'multiples of the median' (MoM) for unaffected (or all) screened in- dividuals at a specified gestational age, so that 1.0 MoM represents the median value ('normal'), 2.0 MoM is twice 'normal', and 0.5 MoM is half 'normal'. The MoM has the advantages that as a ratio it is unitless and so avoids the need to specify the original units of measurement (which vary from centre to centre), that it automatically adjusts for gestation, and that it indicates how high or low a particular value is. In pregnancy, fragments of placental DNA (which reflect the DNA of the fetus) are shed into the maternal circulation and mix with fragments of maternal DNA shed from maternal cells in an approximate ratio of 1:10. DNA analysis, such as counting the number of cell-free DNA fragments that map to chromosome 21, has been used as an antenatal screening test for Down's syndrome, with a high screening perform- ance (detection rate 98–99%, false-positive rate 0.2%), but there is a test failure rate of a few per cent, partly due to a lack of placental DNA in the maternal plasma. At present the test is costly and labour intensive which has tended to preclude its use as a universal screening test. DNA analysis has become part of routine antenatal screening for Down's syndrome and other chromosome disorders (trisomy 18 and 13). A cost-effective screening strategy that has been introduced is antenatal reflex DNA screening in which women who accept screening have a conventional first trimester test. Women are screen-negative unless the first trimester markers yield a risk of having a pregnancy with trisomy 21, 18 or 13 above a certain level (eg ≥ 1 in 800 comprising about 10% of the women) and then automatically have a DNA sequencing ana- lysis using stored plasma from the original blood sample (i.e. a reflex analysis), thereby avoiding the need to recall them for counselling and another blood collection. The conventional test markers and the DNA analysis are considered together as a single test. Neonatal screening Neonatal screening for phenylketonuria, one of the first population- wide screening programmes to be introduced, has proved to be ef- fective and worthwhile in spite of the rarity of the disorder (about 1 per 10 000 births). A low-phenylalanine diet prevents severe mental retard- ation in affected infants. Neonatal screening has prevented cretinism, which is now extremely rare. Additional screening tests could be added to the blood already collected for phenylketonuria and hypothyroidism screening (e.g. MCADD; see Table 2.12.4), and may be justified for other inborn errors of metabolism, given that much of the cost and ef- fort is already spent. However, a line needs to be drawn; tandem mass spectrometry can identify over 40 disorders, but only a handful jus- tify screening as defined. Neonatal screening for congenital deafness is worthwhile and has recently been introduced in the United Kingdom using technology that does not rely on voluntary subject response to noise, thus making it possible to test for hearing deficit in infancy. Screening for congenital dislocation of the hip has been widely prac- tised for many years, without good evidence of efficacy. Galactosaemia (an autosomal recessive inborn error of metabolism) may cause ser- ious illness in the neonate, including septicaemia and encephalop- athy, and cognitive impairment in later life, but it has not been shown that neonatal screening prevents these effects. Neonatal screening for cystic fibrosis has

been introduced in some places without evidence that screening reduces the incidence or severity of the associated lung disease, the main cause of disability and death from cystic fibrosis. Screening in childhood Children are examined routinely to see if they are gaining weight and height as expected and to assess their hearing and vision. There is no evidence that systematic examination of children achieves greater health benefits than encouraging parents to take their child to a doctor if they are concerned, but nonetheless much such activity has taken place. In spite of the lack of formal evidence, it is probably sensible to check the visual acuity of children on starting school, as is current practice in many places. Unfortunately, the lack of evidence to support screening in childhood is often camouflaged in the term 'child- hood surveillance'. As with all screening, evidence of benefit should be sought before acceptance, even if this requires large-scale studies. One disorder that merits screening, is screening for familial hyper- cholesterolaemia, an inherited disorder with a prevalence of about 4 per 1000 that leads to early cardiovascular disease. Parents can be

2.12 Medical screening 147 Table 2.12.4 Summary of neonatal screening tests of proven value Disorder Approximate natural prevalence
(per 10 000 births) in UK Primary screening test Secondary screening test(s) Detection rate (%) False-positive rate (%) Odds of being affected given a positive result Diagnostic test Intervention 1°ry screening test 1°ry and 2°ry screening tests Congenital hypothyroidism 3 T4 or TSH assay before hospital discharge TSH and T4 at 5–7 days 100 20 1:668 1:19 Clinical examination, T4 , free T4 , TSH, thyroid scan Thyroxine Phenylketonuria 1 Serum phenylalanine assay Repeated serum phenylalanine assay 100 0.2 1:22 1:0.05 High plasma phenylalanine (>16.5 mg/dl) using quantitative technique; exclusion of biopterin defects Diet low in phenylalanine Medium chain acyl CoA dehydrogenase deficiency (MCADD) 1 Tandem mass spectrometry (together with PKU) 100 0 1:0 Repeat test Avoidance of fasting, prompt treatment of minor illnesses Deafness 14 Transient evoked otoacoustic emissions (TEORE) Automated auditory brainstem response (AABR) 80 0.6% 1:5 Repeat test Hearing aid or cochlear implant Adapted from Wald NJ, Leck I, eds. Antenatal and Neonatal Screening (2nd ed). (2000) Oxford University Press, Oxford.

148 section 2 Background to medicine screened at the same time in child-parent screening for familial hyper- cholesterolaemia. The method uses the timing of childhood vaccin- ation as a convenient turnstile when a cholesterol measurement is most discriminatory for familial hypercholesterolaemia (1–2 years of age) and a cholesterol measurement used in combination with a DNA mutation analysis. The parents of affected children are offered testing because as the disorder is inherited as an autosomal dominant disorder one parent of an affected child will also be affected. The af- fected parent can be offered preventive statin therapy immediately and the child after the age of ten. The method leads naturally to cas- cade testing in which close relatives are tested: half of all first degree relatives being affected. This protocol has been shown to be feasible and acceptable in a national demonstration project. This method of screening is being considered by various public health agencies. Adult screening Perhaps surprisingly, only a few disorders justify medical screening in adults. These are summarized in Table 2.12.5. Cancers Three cancers meet the screening requirements: breast, cervical, and colorectal. Cervical cancer screening illustrates the principle that effective adult screening programmes require a population age–sex register. Everyone in the appropriate age–sex group for screening can then be identified and sent written invitations at appropriate intervals. Formerly, cervical screening was carried out 'oppor- tunistically' when women happened to consult doctors, and such screening failed because younger

women, at lower risk, see doc- tors more frequently than older women, at higher risk, so cervical smears were carried out on the low-risk group, and at more frequent intervals than necessary for effective screening. It was only with the introduction of a systematic screening programme based on age–sex registers that most older women were screened and cervical cancer mortality fell appreciably in the United Kingdom and other Western countries. Now women are invited 3-yearly between the ages of 25 and 49, and 5-yearly between the ages of 50 and 64. Screening in the United Kingdom is based on testing a cervical brush sample for human papillomavirus (HPV) followed by cytological examination on the same sample if the HPV test is positive (reflex testing). The evidence on efficacy comes from non-randomized studies: screening reduces mortality from cervical cancer by about 80%. Mammographic breast cancer screening is offered in the United Kingdom at 3-yearly intervals to women aged 50 to 70 (though the age range may soon extend down to age 47 and up to 73). Randomized trials have shown that it reduces breast cancer mortality, by about a fifth in a population offered screening or a third in women who accept screening. Manual breast self-examination by women to screen for breast cancer has been shown in randomized trials not to significantly reduce mortality, an observation which illustrates that screening for cancer and other diseases should not be practised simply because it seems intuitively useful: rigorous evidence on efficacy is needed. A colorectal cancer screening programme based on 2-yearly faecal occult blood testing in men and women aged 60 to 70 has been shown in randomized trials to reduce colorectal cancer mortality by about 15% in a population offered screening. Population screening has been introduced in the United Kingdom. A second screening procedure further reduces colorectal cancer mortality. Two randomized trials of once-only flexible sigmoid- oscopy versus no intervention, with identification and resection of colonic polyps, showed a reduction in colorectal cancer mortality of about 30% in a population offered screening, or about 40% in people who attended for screening. There are no published randomized trials of colonoscopy but case–control (observational) studies indi- cate that the mortality reduction is little or no greater than with flex- ible sigmoidoscopy (as fewer cancers occur in the ascending colon and few of the ones that do seem to be prevented). A meta-analysis of cross-sectional studies in people who had both CT colonography and colonoscopy showed that CT detected 48%, 70%, and 84% of all polyps less than 6, 6–9, and more than 9 mm, with a false-positive rate of 7% (i.e. 7% of unaffected people require colonoscopy). This suggests that CT may be an acceptable surrogate for flexible sigmoid- oscopy or colonoscopy: it may be a little less effective in preventing colorectal cancer, but it is less invasive. Chest radiography to screen for lung cancer has been shown in randomized trials not to significantly reduce mortality. However, low-dose spiral computed tomographic screening has been shown in a randomized trial to reduce mortality from lung cancer, by about 20%. The screened group had yearly CT over three years with 6.5 years follow-up; the control group had single view chest X-rays, but since these have been shown to be ineffective the trial can be interpreted as CT versus no screening. However, the false-positive rate was high; about 22% of people without lung cancer had positive screening tests. Much of the diagnostic uncertainty was resolved with further imaging examinations, but about 4% of participants without lung cancer had bronchoscopy, thoracoscopy, or thoracotomy. There is also uncertainty as to whether screening is cost-effective, and con- cern that smokers may derive false reassurance from negative screens and continue to smoke. Launching lung cancer population screening programmes by identifying smokers and offering them a spiral CT examination, while effective, remains an issue for discussion. In systematic population-based cancer screening programmes, only people within a relatively narrow age range are invited for tests (effectively, age is used as the initial screening enquiry). Cancer screening tends to be most effective around the age of 60 in terms of cost per

year of life saved; the lower incidence of cancer in younger people, and the shorter life expectancy in older people, mean that fewer years of life will be gained for the same number of people screened. The justification for a narrow age range is economic. Older people are not turned away, however, and the age range over which women are invited for mammographic screening, for example, widened over time (it was originally 50–64). Usually it is not ap- propriate to stop inviting people for screening examinations above a certain age; if they are fit enough and willing to attend for screening examinations, they are suitable candidates for screening. Cancer screening is generally conducted at 2- to 3-yearly intervals; in prin- ciple, more frequent screening would detect more cancers but the yield per 1000 screening examinations would be lower. Screening for prostate cancer, mainly through measurement of serum prostate-specific antigen (PSA) was introduced into medical practice with no evidence of reduction in mortality. PSA can distin- guish between individuals who will and will not die of prostate cancer. However, discrimination weakens as the interval between the PSA test and clinical presentation or death from the cancer lengthens. By the time the PSA test is highly discriminatory, the disease may be too far advanced for treatment to be effective. The usual cut-off levels pro- posed for PSA screening (c.4 ng/ml) lead to a high proportion of older men being positive. A prostate biopsy in these individuals is often positive, because 25% of prostates in men aged 70 have histological

2.12 Medical screening 149 Table 2.12.5 Summary of adult screening for selected disorders Disorder Prevalence Screening procedure Age range Subsequent investigation Detection rate Positive rate Odds of disorder in screen positives Uptake of screening Treatment Reduction in disease Breast cancer 4% of all deaths (women) Mammography 2–3-yearly 50+ Further imaging; fine needle biopsy Not applicable 8% first screen, 4% subsequent; biopsy rate 0.8% 1:6 (2:1 among women biopsied) 70–80% Surgery (± chemotherapy, radiotherapy) 24% reduction in mortality at age 50–74; 16% at age 40–49 (from meta-analyses of randomized trials) Colorectal cancer 3% of all deaths (men and women) Faecal occult blood testing 2-yearly 60+ Colonoscopy ± barium enema Not applicable 2–3% 1:10 50–60% Surgery 15–18% reduction in mortality (from two randomized trials) Cervical cancer 0.5% of all deaths (women) Cervical smear ± HPV testing 3–5 yearly 25+ Repeat smear in 6 months (mild dyskaryosis); colposcopy (moderate/severe dyskaryosis) Not applicable 5–10% (higher in younger than older women), lower with HPV test and smear – 80% Local ablation or excision (rarely hysterectomy) 90% reduction in mortality (from case–control studies) Diabetic retinopathy Proliferative retinopathy 50% IDDM 50% NIDDM Macular oedema 15% IDDM 10% NIDDM Retinal photography with mydriatic yearly All Assessment by ophthalmologist 78% 0 50% if done in hospital clinics Photocoagulation Reduction in blindness >90% (proliferative retinopathy) 65% (macular oedema) Abdominal aortic aneurysm rupture Men aged 65+ 2% of all deaths 7% have aortic diameter ≥3.0 cm Ultrasound scan 65 (men) CT or MRI 86% 0.6% 75% Open surgery Chlamydia trachomatis genital infection (subsequently causing PID) Chlamydia: 5% among women under 25 PID 2% Nucleic acid amplification test on urine sample <25 (sexually active) – 90–95% <1% 64% Doxycycline or azithromycin 56% reduction in PID HPV, human papillomavirus; IDDM, type 1 diabetes; NIDDM, type 2 diabetes; PID, pelvic inflammatory disease (Ashton et al., 2002).

150 section 2 Background to medicine evidence of cancer even though only a small minority of these men will suffer from or die of the disease. Such ‘overdiagnosis’ (the diag- nosis of cancers that would otherwise never have come to clinical at- tention) is a potentially serious problem in cancer screening. These cancers are best never diagnosed; once diagnosed, anxiety and un-

necessary hazardous investigation and treatment will ensue. In 2009, two randomized trials of PSA screening for prostate cancer were reported, one showing a significant ($p = 0.01$) 20% re- duction in prostate cancer mortality in men invited for screening (27% in those who were screened) and the other one showing a nonsignificant increase, but consistent with a 15% reduction. Both trials showed a high rate of overdiagnosis; in the larger of the two trials, for every one prostate cancer death prevented, 1410 men were screened, of whom 16% (230) had a biopsy, identifying 49 prostate cancers of which 48 were treated unnecessarily. Taking the two trials together, screening for prostate cancer by PSA testing probably does reduce prostate cancer mortality. Subsequent smaller trials have shown similar results. The reduction, however, may not be judged sufficient to warrant the level of overdiagnosis which leads to many men receiving unnecessary hazardous treatment. A randomized trial of ovarian cancer screening using ultrasound examination of the ovaries and measurement of a serum protein marker (CA125), found no statistically significant difference be- tween women randomized to screening and those randomized to ‘no screening’. While the efficacy of such screening cannot be com- pletely excluded the evidence suggests that, if there is an effect, it is small. A cancer screening programme that is currently under inves- tigation is screening for future stomach cancer by identifying people with Helicobacter pylori infection of the stomach. Nonmalignant diseases Screening for abdominal aortic aneurysms that, in the absence of surgery, are likely to rupture, by the ultrasound measurement of the aortic diameter, is worthwhile. The test is very discriminatory (see Fig. 2.12.6). Ruptured abdominal aortic aneurysms account for 2% of all deaths in men over 65, but are rare when the maximal aortic diameter is less than 5 cm. In the United Kingdom a screening pro- gramme based on abdominal aortic diameter using ultrasound is in progress for men aged 65 (rupture is rare in younger men). Over all ages ruptured abdominal aortic aneurysm is about twice as common in men as in women. Mortality rates for women for women are similar to those in men about 10 years younger. Most men will need only a single scan in the year in which they reach 65. Screening people with diabetes for retinopathy using retinal pho- tography is very effective; it has been shown in randomized trials to reduce blindness by 90% with proliferative retinopathy and 65% with macular oedema. A national screening programme operates in the United Kingdom, based on inviting people from diabetic regis- ters held in general practice. Chlamydia infection in young women causes pelvic inflammatory disease (which may be complicated by chronic pelvic pain, ectopic pregnancy, and tubal infertility and, when giving birth, causes neo- natal eye and lung damage). Screening for chlamydia infection based on urine samples is followed by short term antibiotic treatment and is effective. Screening women under 25 has been recommended but no systematic screening programme has been introduced in the United Kingdom. Much screening activity falls under the category of ‘risk factor screening’ and such screening tends to be ineffective (e.g. cholesterol testing in screening for future ischaemic heart disease events; see Fig. 2.12.7), blood pressure measurement in screening for future stroke, and bone density measurement as a screening test for future hip fractures. The problem arises because, for the reasons given here, risk factors that may be important causes of disease are usually poor screening tests. Most adults have high serum cholesterol and high blood pressure relative to levels in young adults (say at age 20), and all postmenopausal women have low bone density relative to pre- menopausal women, so nearly all older adults are ‘exposed’. Fig. 2.12.8 shows the effect of combining different markers on the detection rate and false-positive rate where several markers that each have a detection rate for a 5% false-positive rate (DR5) of 10%, 15%, or 20% and the standard deviation is the same in affected and unaffected individuals. Only when tests individually have a DR5 of about 20% or greater will multiple marker screening become a real- istic proposition. For example, combining five

relatively weak inde- pendent markers, each with a DR5 of 15%, yields only a 40% overall detection rate for a 5% false-positive rate, and combining ten yields a 60% detection rate for the same 5% false-positive rate. At present, screening for future coronary disease and most other diseases using causal risk factors is not effective because even in combination they are not sufficiently discriminatory. Hypothyroidism in adults is widely regarded as a preventable cause of lethargy and depression. This has prompted attempts at screening for this disorder by measuring levels of thyroxine (T4) or thyroid-stimulating hormone (TSH) and classifying individ- uals as positive if TSH is above or T4 below the relevant reference range (which is usually the 95th centile range in the population). This is an example of the ‘tautological screening’ that arises from defining a disorder in terms of the test used to screen for it (see earlier). The solution to this circularity is to identify individuals from a population with TSH or T4 outside specified TSH or T4 limits and then offer each, in random order, thyroxine or placebo to determine whether thyroxine treatment relieves the symptoms more often than can be explained by chance. Each person is therefore their own control, and the response to treatment defines the clinical disorder. Such a cross-over randomized trial has been 100% 80% Number of screening markers 0 5 Detection rate for multiple markers 10 15 20 25 30 35 40 60% 40% 20% 0% Fig. 2.12.8 Overall screening performance from combining individual screening markers: detection rate for a 5% false-positive rate (DR5) according to the number of screening markers combined that individually have a DR5 of 10% or 15% or 20%. Reproduced from Wald N, Morris J and Rish S, ‘The efficacy of combining several risk factors as a screening test’, J Med Screen 2005; 12: 197–201. London: Royal Society of Medicine Press, 2005.

 done and showed that screening for hypothyroidism is worthwhile. In screening the same approach should be used to identify which individuals will benefit from treatment. Clarity of terminology and purpose Certain terms used in screening are probably best avoided because they lack clarity. The term ‘carrier screening’ implies that carriers of autosomal recessive disorders (e.g. cystic fibrosis) themselves have a disease; they do not. The goal of such screening is to identify couples who are both carriers. ‘Couple screening’ involves collecting samples from both parents and reporting a positive result only when both are carriers. The term ‘genetic screening’ lacks clarity and tends to imply screening for inherited disorders even though some genetic disorders that are screened for (e.g. Down’s syndrome) are usually not inherited. The term creates a false impression that something special is being offered that other forms of screening lack. For many people genes and consequently all things genetic are seen as highly determinant, even inevitable, influences, which is usually not the case. Genetic markers of a disease are in most instances too insensitive and too nonspecific for screening purposes. The term ‘case finding’ often implies the identifica- tion of cases of the disorder being screened for, while in fact it identifies individuals with a positive screening test for that disorder. For example, a case of ‘hypertension’ relates to the test result (high blood pressure), not the diseases it causes. The term ‘opportunistic screening’ is a eu- phemism for nonsystematic and nonorganized screening. The purpose of medical screening is clear—to avoid disability and premature death at an acceptable level of safety. Determining efficacy is essential. Many screening tests are effective and should be part of public health practice. But particular care is needed in evaluating tests that arise out of technological development in the absence of a clear case of medical need. Whole-body scanning using MRI and fetal ultra- sound examination are examples. Such screening, without defining the specific disorders being screened for, detects ‘incidentalomas’ (so- called ‘abnormal’ findings with little or no knowledge of their med- ical significance). There is no place for such screening in responsible medical practice. For example, total body MRI scanning is now ad- vertised to the public as a screening test with little attention paid to whether it prevents serious disability or

death, or meets the criteria set out in Box 2.12.1. A routine fetal anomalies scan at about 18 weeks of pregnancy has some proven specific applications (e.g. the detection of anencephaly, severe congenital heart disease, and placenta praevia ex- tending to cover the internal cervical os), but the term ‘fetal anomaly screening’ lacks specificity. The challenge in performing a scan is to seek these specific anomalies but not to report other ‘incidentalomas’ which will undoubtedly lead to parental anxiety and further investigation but for which early detection has not been shown to be worthwhile. Under the ambiguous heading of genetic screening, so-called ‘gene chips’ have been developed, that can detect in one test many hundreds of genetic mutations with little or no evidence that knowledge of these will lead to useful medical intervention that will improve the health and quality of lives of the people being so tested. Medical screening needs to be driven by the medical need, not the technological capacity. Doctors have a professional responsibility to discourage techno- logically driven screening and to ensure that all screening meets the requirements set out in Box 2.12.1. Screening promoted only in terms of the application of a particular technology should not be part of medical practice. FURTHER READING Abu-Helalah M, Law MR, Bestwick JP, Monson JP, Wald NJ (2011). A randomized double blind cross-over trial to investigate the efficacy of screening for adult hypothyroidism. J Med Screen, 17, 164–9. Ashton HA, et al. (2002). The Multicentre Aneurysm Screening Study (MASS) into the effect of abdominal aortic aneurysm screening on mortality in men: a randomised controlled trial. Lancet, 360, 1531–9. Atkins WS, et al. (2010). Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised con- trolled trial. Lancet, 375, 1624–33. Breslow N, et al. (1977). Latent carcinoma of prostate at autopsy in seven areas. Int J Cancer, 20, 680–88. Holland WW, Stewart S (2005). Screening in disease prevention. Nuffield Trust, Radcliffe Publishing, Oxford. Journal of Medical Screening: screening briefs (1994a, 1994b, 1995, 1996, 1997). J Med Screen, 1, 73; 1, 255; 2, 126; 3, 110; 4, 54. Law MR, Morris J, Wald NJ (1994). Screening for abdominal aortic aneurysms. J Med Screen, 1, 110–16. McKeown T (1968). Validation of screening procedures. In: Screening in medical care. Reviewing the evidence. Nuffield Provincial Hospital Trust, Oxford University Press, Oxford. Mulhall BP, Veerappan GR, Jackson JL (2005). Meta-analysis: com- puted tomographic colonography. Ann Intern Med, 142, 635–50. National Lung Screening Trial Research Team (2011). Reduced lung- cancer mortality with low-dose computed tomographic screening. N Engl J Med, 365, 395–409. Palomaki GE, et al. (2012). DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. Genet Med, 14, 296–305. Scoen RE, et al. (2012). Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy. N Engl J Med, 366, 2345–57. Thorner RM, Remein QR (1961). Principles and procedures in the eval- uation of screening for disease. Public Health Monograph No 67. Public Health Service Publication No 846. US Department of Health Education and Welfare, Washington, DC. Wald DS, Bestwick JP, Morris JK, Whyte K, Jenkins L, Wald NJ (2016). Child-parent familial hypercholesterolemia screening in primary care. N Engl J Med, 375, 1628–37. Wald NJ (1994). Guidance on terminology. J Med Screen, 1, 76. Wald NJ (2004). The epidemiological approach: an introduction to epi- demiology in medicine, 4th edition. Wolfson Institute of Preventive Medicine/Royal Society of Medicine Press, London. Wald NJ, Cuckle H (1989). Reporting the assessment of screening and diagnostic tests. Br J Obstet Gynecol, 96, 389–96. Wald NJ, Hackshaw AK, Frost CD (1999). When can a risk factor be used as a worthwhile screening test? BMJ, 319, 1562–5. Wald NJ, Leck I (eds) (2000). Antenatal and neonatal screening, 2nd edition. Oxford University Press, Oxford. Wald NJ, Law MR (2003). A strategy to reduce cardiovascular disease by more than 80%. BMJ, 326, 1419–23. Wald NJ, et al. (2004). SURUSS in perspective. Br J Obstet Gynaecol, 111, 521–31. Wald NJ, Morris JK, Rish S (2005). The efficacy of combining several risk factors as a screening test. J Med Screen, 12, 197–201. Wald NJ, et al. (2018). Prenatal reflex

DNA screening for trisomies 21, 18, and 13. Genet Med, 20, 825–30. Wilson JMS, Jungner G (1968). Principles and practice of screening for disease. WHO Public Health Paper No. 34, World Health Organization, Geneva.

Revision #1
Created 2026-01-22 16:40:54 UTC by Omar Ayman
Updated 2026-01-22 16:40:55 UTC by Omar Ayman