4. The Phenomenon of Disease

Concepts in defining, classifying, detecting, and tracking disease and other health states. The concept of natural history – the spectrum of development and manifestations of pathological conditions in individuals and populations.

Definition and classification of disease

Although the public health profession is sometimes inclined to refer to the health care system as a "disease care system", others have observed that public health also tends to be preoccupied with disease. One problem with these charges is that both "health" and "disease" are elusive concepts.

Defining health and disease

Rene Dubos (*Man Adapting*, p348) derided dictionaries and encyclopedias of the mid-20th century for defining "disease as any departure from the state of health and health as a state of normalcy free from disease or pain". In their use of the terms "normal" and "pathological", contemporary definitions (see table) have not entirely avoided an element of circularity.

Rejecting the possibility of defining health and disease in the abstract, Dubos saw the criteria for health as conditioned by the social norms, history, aspirations, values, and the environment, a perspective that remains the case today (Temple *et al.*, 2001). Thus diseases that are very widespread may come to be considered as "normal" or an inevitable part of life. Dubos observed that in a certain South American tribe, pinta (dyschromic spirochetosis) was so common that the Indians regarded those *without* it as being ill. Japanese physicians have regarded chronic bronchitis and asthma as unavoidable complaints, and in the mid-19th century U.S., Lemuel Shattuck wrote that tuberculosis created little alarm because of its constant presence (Dubos, 251). As for the idealistic vision of health embodied in the WHO Constitution, Dubos wrote:

> ". . . positive health . . . is only a mirage, because man in the real world must face the physical, biological, and social forces of his environment, which are forever changing, usually in an unpredictable manner, and frequently with dangerous consequences for him as a person and for the human species in general." (*Man Adapting*, 349)

With the sequencing of the human genome, the question of what is disease arises must be dealt with lest every genetic variation or abnormality be labeled as disease-associated (Temple *et al.*, 2001). Such labeling can have severe ramifications or alternatively be beneficial. Temple *et al.* reject Boorse's definition ["a type of internal state which is either an impairment of normal functional ability – that is, a reduction of one or more functional abilities below typical efficiency – or a

limiation on functional ability caused by environmental agents"[*]] as clinically impractical and not helpful for simplifying interpretation of genetic variations. These authors assert that the key element is risk of adverse consequences and offer the definition "disease is a *state* that places individuals at *increased risk of adverse consequences*" (Temple *et al.*, 2001, p807, italics in original). The World Health Organization classifies adverse consequences as including physical or psychological impairment, activity restrictions, and/or role limitations, though these may be culturally-dependent (Temple *et al.*, 2001, p808). Indeed, since the risk of adverse consequences is often variable across patients, Temple *et al.* suggest that the "'cutoff' between the categories of diseased and nondiseased could be based on many factors, including … potential for treatment" (p808) and that if the risk from a genetic abnormality is very low it may be better characterized as a "risk factor" than a "disease". In response to a criticism from Gerald Byrne (*Science* 7 Sept 2001;293:1765-1766), James Wright (a co-author of Temple *et al.*) acknowledges that no definition will work in all contexts, offers yet another definition for dealing with risk-taking behaviors, and suggests that "given the potential genetic explanations for behavioral disorders (2), with time … mountain climbing might be viewed by some as [a disease manifestation]" (p.1766; reference 2 is a paper by DE Comings and K Blum in *Prog Brain Res* 2000)!

Clearly, general definitions of health and disease involve biological, sociological, political, philosophical, and many other considerations. Such definitions also have important implications, since they delimit appropriate arenas for epidemiology and public health. But even with a consensus on a general definition, we will still face major challenges in recognizing and classifying the myriad diversity of health-related phenomena encountered in epidemiology and other health (disease) sciences.

Some definitions of disease and health

Dorland's Illustrated Medical Dictionary (28th ed., Phila, Saunders, 1994):

Disease – "any deviation from or interruption of the normal structure or function of any part, organ, or system (or combination thereof) of the body that is manifested by a characteristic set of symptoms and signs . . .".

Health – "a state of optimal physical, mental, and social well-being, and not merely the absence of disease and infirmity."

Stedman's Medical Dictionary (26th ed., Baltimore, Williams & Wilkins, 1995):

Disease –

1. An interruption, cessation, or disorder of body functions, systems, or organs;

[*] C. Boorse, in *What is disease*? In: Humber JM, RF Almeder, eds, Biomedical ethics reviews, Humana Press, Totowo NJ, 1997, pp.7-8, quoted in Temple *et al.* (2001), p807.

2. A morbid entity characterized usually by at least two of these criteria: recognized etiologic agent(s), identifiable group of signs and symptoms, or consistent anatomical alterations.
3. Literally dis-ease, the opposite of ease, when something is wrong with a bodily function."

Health

1. The state of the organism when it functions optimally without evidence of disease or abnormality.
2. A state of dynamic balance in which an individual's or a group's capacity to cope with all the circumstances of living is at an optimum level.
3. A state characterized by anatomical, physiological, and psychological integrity; ability to perform personally valued family, work, and community roles; ability to deal with physical, biological, and psychological and social stress; a feeling of well-being; freedom from the risk of disease and untimely death."

Taber's Cyclopedic Medical Dictionary (17th ed. Phila., FA Davis, 1993. Ed. Clayton L. Thomas):

Disease – "Literally the lack of ease; a pathological condition of the body that presents a group of clinical signs and symptoms and laboratory findings peculiar to it and that sets the condition apart as an abnormal entity differing from other normal or pathological body states. The concept of disease may include the condition of illness or suffering not necessarily arising from pathological changes in the body. There is a major distinction between disease and illness in that the former is usually tangible and may even be measured, whereas illness is highly individual and personal, as with pain, suffering, and distress." [Examples given include: hypertension is a disease but not an illness; hysteria or mental illness are illnesses but have no evidence of disease as measured by pathological changes in the body.]

Classification is the foundation

As stated in an early (1957) edition of the *Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death* (ICD):

> "Classification is fundamental to the quantitative study of any phenomenon. It is recognized as the basis of all scientific generalization and is therefore an essential element in statistical methodology. Uniform definitions and uniform systems of classification are prerequisites in the advancement of scientific knowledge. In the study of illness and death, therefore, a standard classification of disease and injury for statistical purposes is essential." (Introduction, pp. vii-ix)

The eminent Australian statistician, Sir George H. Knibbs, credited Francois Bossier de Lacroix (1706-1777), better known as Sauvages, with the first attempt to classify diseases systematically, in his *Nosologia Methodica.*

A classification is not merely a set of names to be applied to phenomena, although a ***nomenclature*** – a list or catalog of approved terms for describing and recording observations – is essential. As explained in the ICD:

> "Any morbid condition that can be specifically described will need a specific designation in a nomenclature. . . This complete specificity of a nomenclature prevents it from serving satisfactorily as a statistical classification [which focuses on groups of cases and aims to bring together those cases that have considerable affinity]. . . . A statistical classification of disease must be confined to a limited number of categories which will encompass the entire range of morbid conditions. The categories should be chosen so that they will facilitate the statistical study of disease phenomena.
>
> "Before a statistical classification can be put into actual use, it is necessary that a decision be reached as to the inclusions for each category. . . . If medical nomenclature were uniform and standard, such a task would be simple and quite direct. Actually the doctors who practise and who will be making entries in medical records or writing medical certificates of death were educated at different medical schools and over a period of more than fifty years. As a result, the medical entries on sickness records, hospital records, and death certificates are certain to be of mixed terminology which cannot be modernized or standardized by the wave of any magician's wand. All these terms, good and bad, must be provided for as inclusions in a statistical classification."

There is not necessarily a "correct" classification system. In classifying disease conditions, choices and compromises must be made among classifications based on etiology, anatomical site, age, and circumstance of onset, as well as on the quality of information available on medical reports. There may also need to be adjustments to meet varied requirements of vital statistics offices, hospitals, armed forces medical services, social insurance organizations, sickness surveys, and numerous other agencies. The suitability of a particular system depends in part on the use to be made of the statistics compiled and in part on the information available in deriving and applying the system.

Defining and measuring the phenomena

Perhaps the first and most important issue in planning or interpreting an epidemiologic study is the definition and measurement of the disease and/or phenomena under study. How satisfactorily this issue can be addressed depends on the nature of the phenomena, the extent of knowledge about it, and the capability of available technology. The specific circumstances can range from the report of a case or series of cases that do not fit the characteristics of any known disorder to a disease that has been thoroughly studied and for which highly accurate and specific diagnostic procedures are available.

In the former category would fall the investigation of the condition that now bears the label chronic fatigue syndrome, where a vague collection of nonspecific symptoms was proposed to constitute a previously unrecognized disease entity, which still awaits a consensus regarding its existence. In situations such as these, a first task is formulating at least a provisional case definition in order to

proceed with the investigation. In the latter category would fall rabies, where a specific, highly virulent organism has been identified and produces characteristic manifestations. Psychiatric disorders would fall somewhere in between. The nub of the problem is that the clarity with which features of the condition – etiologic factors, co-factors, natural history, response to treatment – can be linked to it depends on how effective are definition and measurement at excluding other entities whose different features will become mixed with those truly characteristic of the condition.

Consider an example. Although angina pectoris had been described in the 18th century (by William Heberden), and some 19th century physicians recognized an association between this symptom and coronary artery sclerosis found at autopsy, the syndrome of acute myocardial infarction (MI) was not recognized until the 20th century. According to W. Bruce Fye [The delayed diagnosis of myocardial infarction: it took half a century. *Circulation* 1985; 72:262-271)] the delay was due to the belief until 1912 that MI was invariably fatal and also to (1) the inconstant relationship of symptoms to pathological findings, (2) excessive reliance on auscultation as an indicator of cardiac disease, (3) failure to routinely examine coronary arteries or myocardium at autopsy, (4) tardiness of clinicians to incorporate new pathophysiologic discoveries into medical practice, (5) willingness to accept theories of disease not supported by scientific evidence, (6) pre-occupation with the new field of bacteriology, and (7) the lack of diagnostic techniques with which to objectively identify coronary artery obstruction or its consequences during life. (This list of reasons fits very well into Thomas Kuhn's description of the process of paradigm shifts – see citation in chapters 1 and 2.)

Classification criteria and disease definition

Since no two entities are completely identical, we (often unconsciously) group them together or differentiate between them according to what we believe to be important for our purposes. Even conditions with different etiologies may nevertheless have the same prognosis or the same response to treatment. Decisions about how far to subdivide categories of what appears to be a single entity depend, therefore, on the difference it may make, the level of knowledge, and our conceptual model.

As we gain more sophisticated understanding of the pathophysiological and biochemical mechanisms of disease conditions – to which the dramatic advances in molecular biology have contributed greatly – opportunities to differentiate among conditions now treated as a single entity and questions about whether to do so are becoming more frequent. For example, a mutation in the p53 gene is present in about 50% of cancers. Should cancers be classified according to whether or not an aberrant p53 gene is present? Is this aspect more important than the anatomical site or the histologic type? If two cancers of the same site and histologic type have mutations at different loci of p53, should they be classified apart?

There are two broad approaches to defining a disease entity. These two approaches are manifestational criteria and causal criteria [see discussion in MacMahon and Pugh].

Manifestional criteria

Manifestational criteria refer to symptoms, signs, behavior, laboratory findings, onset, course, prognosis, response to treatment, and other manifestations of the condition. Defining a disease in terms of manifestational criteria relies on the proposition that diseases have a characteristic set of

manifestations. The term "syndrome" (literally, "running together" [Feinstein, 2001]) is often applied to a group of symptoms or other manifestations that apparently represent a disease or condition whose etiology is as yet unknown. Most chronic and psychiatric diseases are defined by manifestational criteria (examples: diabetes mellitus, schizophrenia, cancers, coronary heart disease).

Causal criteria

Causal criteria refer to the etiology of the condition, which, of course, must have been identified in order to employ them. Causal criteria are most readily available when the condition is simply defined as the consequences of a given agent or process (e.g., birth trauma, lead poisoning). The other group of conditions where causal criteria are available consists mostly of infectious diseases for which the pathogen is known, e.g., measles. Through the use of causal criteria, diverse manifestations recognized as arising from the same etiologic agent (e.g., the various presentations of infection with *Treponema pallidum* [syphilis] or with *Borrelia burgdorferi* [Lyme disease]) can be classified as the same disease entity. Similarly, conditions that have a similar presentation (e.g., gonorrhea, chlamydia) can be differentiated. Temple *et al.* (2001) associate these two approaches with two opposing schools, which they term, respectively, "nominalist" (defining disease in terms of labeling symptoms) and "essentialist (reductionist)" (defining disease in terms of underlying pathological etiology). [Scadding suggests that the nominalist approach may be "roughly accurate", whereas the essentialist approach may be "precisely wrong"*.]

Manifestational versus causal criteria

The rationale for defining diseases based on manifestational criteria is borne largely of necessity – until we know the etiology, what else can we do? – and partly of the expectation that conditions with similar manifestations are likely to have the same or at least related etiology. Although this expectation has often been fulfilled, it is by no means a certainty. Simply because two conditions have identical manifestations (to the extent that we are currently able to and knowledgeable enough to measure these) does not ensure that they are the same entity in all other relevant respects, notably etiology. For example, even if major depressive disorder could be diagnosed with 100% agreement among psychiatric experts, the possibility would still exist that the diagnosis embraces multiple disease entities with very different etiologies. Similarly, an etiologic process that leads to major depressive disorder may be expressed with different manifestations depending upon circumstances and host characteristics.

Replacement of manifestional criteria by causal criteria

Nevertheless, the process seems to work. The evolution of the definition and detection of a disease, with the replacement of definitions based on manifestational criteria with definitions based on causal criteria, is well illustrated by HIV/AIDS. In 1981, clinicians in San Francisco reported seeing young American men with Kaposi's sarcoma, a tumor previously seen only in elderly Mediterranean males. Physicians in Europe found similar tumors in people from Africa. Shortly afterward, the Centers for Disease Control (CDC) noted that requests for pentamidine, a rarely prescribed antibiotic used for

* Scadding G, *Lancet* 1996;348:594, cited in Temple *et al.* (2001), p808.

treating pneumocystis carinii pneumonia (PCP – an opportunistic infection generally seen only in medically immunosuppressed patients), had increased sharply from California. Investigation revealed that PCP was occurring in apparently otherwise healthy young men.

A first step in investigating a new, or at least different, disease is to formulate a case definition that can serve as the basis for identifying cases and conducting surveillance. The Acquired Immunodeficiency Syndrome (AIDS) was defined by the CDC in terms of manifestational criteria as a basis for instituting surveillance (reporting and tracking) of this apparently new disease. The operational definition grouped diverse manifestations – Kaposi's sarcoma outside its usual subpopulation, PCP and other opportunistic infections in people with no known basis for immunodeficiency – into a single entity on the basis of similar epidemiologic observations (similar population affected, similar geographical distribution) and their sharing of a particular type of immunity deficit (elevated ratio of T-suppressor to T-helper lymphocytes).

After several years human immunodeficiency virus (HIV, previously called human lymphotrophic virus type III) was discovered and demonstrated to be the causal agent for AIDS, so that AIDS could then be defined by causal criteria. However, because of the long latency between infection and the development of AIDS, manifestational criteria are still a part of the definition of AIDS, though not of HIV infection itself. The original CDC reporting definition was modified in 1985 (*Morbidity and Mortality Weekly Report* [*MMWR*] 1985;34:373-5) and again in 1987 (*MMWR* 1987:36 [suppl. no. 1S]:1S-15S) to incorporate (1) a broader range of AIDS-indicator diseases and conditions and (2) HIV diagnostic tests. The proportions of AIDS cases that meet only the newer definitions vary by gender, race, and risk category.

In parallel with the institution of U.S. reporting definitions there has been an evolution in the international disease classification for AIDS. An original interim ICD classification for AIDS was issued on October 1, 1986, with the expectation that periodic revisions would be required. The first revision (January 1, 1988) characterized the causal agent and the change in terminology from human T-cell lymphotropic virus-III (HTLV-III) to HIV (Centers for Disease Control. Human immunodeficiency virus (HIV) infection codes and new codes for Kaposi's sarcoma: official authorized addenda ICD-9-CM (Revision 2) effective October 1,1991. *MMWR* 1991; 40(RR-9):1-19). The 1991 revision dealt only with morbidity reporting and provided for more detail about manifestations of HIV infection. All manifestations of HIV infection were to be coded, but a hierarchical classification was made for the stage of HIV infection. Distinctions were made between conditions occurring with HIV infection (e.g., 042.0: HIV with toxoplasmosis) and those occurring due to HIV infection (e.g., 042.1: HIV causing tuberculosis).

To recapitulate the above discussion, where we are fortunate, the classification based on manifestational criteria will closely correspond with that based on causal criteria but this is by no means assured because:

1. A single causal agent may have polymorphous effects (e.g., cigarette smoking is a causal factor for diverse diseases, herpes zoster causes chicken pox and shingles);

2. Multiple etiologic pathways may lead to identical (or at least apparently identical) manifestations, so that a (manifestationally-defined) disease entity may include subgroups with differing etiologies;
3. Multicausation necessitates a degree of arbitrariness in assigning a single or necessary cause to a disease category. For example, nutritional status and genetic constitution are contributing factors for tuberculosis. Had medical knowledge developed differently, tuberculosis might be known as a nutritional disorder with the bacillus as a contributory factor.
4. Often, not all persons with the causal agent (e.g., hepatitis A) develop the disease.

In actual epidemiologic practice, most disease definitions are based on manifestational criteria and proceed on the general assumption that the greater the similarity of the manifestations, the more likely the illness represents a unitary disease entity. The objective is to form classifications that will be useful in terms of studying the natural history of the disease and its etiology and also for treatment and prevention. There are numerous contemporary (e.g., Gulf War syndrome, chronic fatigue syndrome) as well as historical examples of this basic approach. In his essay on "The Blame-X syndrome", Feinstein (2001) points to some of the difficulties that arise in linking manifestations to etiology when a causative pathophysiologic processes has not been identified and how cultural, social, political, and legal factors become bound up with the scientific questions.

Disease classification systems

As diseases are defined they are organized into a classification. The primary disease classification system in use is the *International Classification of Disease* (ICD), now published by the World Health Organization. Introduced in 1900 for the purposes of classifying causes of death, the ICD apparently has its origins in a list of categories prepared by William Farr and Marc D'Espine in 1853 (see Feinstein, 2001, for citation). The ICD was expanded to cover illness and injury in 1948. In the United States, the National Center for Health Statistics publishes an adapted version of the ICD to incorporate syndromes and illnesses not listed in the WHO edition. The American Psychiatric Association performs a similar function for classification of mental disorders, with its *Diagnostic and Statistics Manual* (DSM) (see below).

Disease classification systems do not necessarily provide the kind of information needed for public health research and policymaking. Diseases and deaths related to tobacco use, for example, cannot be identified from ICD codes, though there has been a movement toward having tobacco use appear as a cause on the death certificate. In the injury area, injuries are classified according to the nature of the injury (e.g., laceration, puncture, burn) rather than the nature of the force that caused it (e.g., gunshot, fire, automobile crash, fall). Injury prevention researchers advocate the use of E (External) codes to permit tabulation by the external cause of the injury.

Classification systems, of course, must be periodically revised to conform to new knowledge and re-conceptualizations. Revisions typically include changes in:

1. usage of diagnostic terms (e.g., for heart disease);
2. disease definitions;

3. organization of categories based on new perceptions about similarities among conditions (e.g., joint occurrence of hypertension and CHD);
4. coding rule (e.g., priorities for selecting an underlying cause of death when multiple diseases are present).

Such changes come at a price, in the form of discontinuities in disease rates over time and confusion for the unwary. For example, prior to 1986, carcinoid tumors were reportable to the National Cancer Institute's Surveillance, Epidemiology, and End Results program (SEER) only if they were specifically described as "malignant gastric carcinoid." In 1986, any tumor described as "gastric carcinoid" was considered malignant and therefore was reportable to SEER. This change produced a substantial rise in the rate of gastric carcinoid tumors in 1986.

Similar problems in comparing rates over time, across geographical area, or among different health care providers can arise from differences or changes in "diagnostic custom" or terminology preferences (see Sorlie and Gold, 1987). In addition, the Diagnosis Related Group (DRG) system introduced in the United States to control the costs of federal reimbursement to hospitals for health care has undoubtedly influenced discharge diagnoses in favor of those with higher reimbursement opportunity. See Feinstein (2001) for more on these and other issues of nomenclature and classification.

Conceptual questions in classifying diseases

Even without the complicating factors of diagnostic custom or changes in classification systems, by its very nature classification poses difficult conceptual questions whose resolutions underlie the disease definitions we employ. Some examples:

1. What constitutes "similarity"?
   Examples: adult versus juvenile onset diabetes; melanoma in the retina versus in the skin; pneumonia of viral, bacterial, or chemical origin; cancers with different genetic "signatures".
2. What is the appropriate cutpoint on a continuum?
   Examples: blood pressure and hypertension; plasma glucose and diabetes; alcohol consumption and alcoholism; fetal death and gestational age.
3. How should ambiguous situations be handled?
   Examples: hypertension controlled with drugs; subclinical infection; alcoholism, schizophrenia or depressive disorder in remission.

As perceptions and understanding changes, so do the answers to the questions. For example, in moving from DSM-III to DSM-IV, the American Psychiatric Association removed the distinction between "organic" and "inorganic" psychiatric disorders, added categories for premenstrual syndrome and gender identity problems, and introduced a V-code (nonpathological descriptor) for religious or spiritual problems ("Psychiatrists set to approve DSM-IV", *JAMA* 7/7/93, 270(1):13-15).

Classifying cause of death

Since mortality data are generally the most widely available, epidemiologists encounter the above problems most often in evaluating the accuracy of cause-specific mortality rates. Cause-specific mortality rates are tabulated using the "underlying cause of death", and until recently this was the only cause available in electronic form. The underlying cause of death is defined as "the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances or violence which preceded the fatal injury" (*Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death*, Geneva, WHO, 1977: 609-701, quoted in Kircher and Anderson, *JAMA* 1987:349). According to Kircher and Anderson, most physicians confuse cause and mechanism. For example, congestive heart failure, cardiorespiratory arrest, asphyxia, renal failure are mechanisms – the means by which the cause exerts its lethal effect.

The following are additional operational problems in assigning a cause of death (see Israel *et al.* 1986):

1. Many conditions can coexist without a direct etiologic chain. When a combination of causes is forced into a single cause, the choice may be arbitrary, even if systematic, and the true circumstances obscured.
2. There is confusion about disease terms; e.g., it is often unclear whether "metastatic" disease refers to a primary or secondary tumor.
3. There is confusion among certifiers about the meaning of classification terms (e.g., "underlying", "immediate", and "contributory" causes). [Confusion is perhaps to be expected, given the complexity of concept and circumstances. According to the ICD, "The words 'due to (or as a consequence of)' . . . include not only etiological or pathological sequences, but also sequences where there is no such direct causation but where an antecedent condition is believed to have prepared the way for the direct cause by damage to tissues or impairment of function even after a long interval." (*Manual of the international statistical classification of diseases, injuries, and causes of death, based on the recommendations of the Ninth Revision Conference*, 1975. Geneva: WHO, 1977:700, quoted in *MMWR* 1991 (26 July);40:3)]
4. Death certificates are often completed late at night or in haste, sometimes to speed funeral arrangements, by a sleepless physician who has never seen the deceased before and for whom care of the living is understandably a higher priority. Partly for these reasons death certificate information is often sloppy or incomplete. Amended certificates with more complete information can be but are rarely filed, and unlikely diagnoses are rarely queried.

Both mortality statistics and case ascertainment for epidemiologic studies can readily be affected by such problems and circumstances (see Percy, *et al.* 1981). Epidemiologic studies for which cause of death is important often have a copy of each death certificate reviewed by a trained nosologist, an expert in classifying diseases, to confirm or correct questionable cause of death information. If resources are available, medical records may be obtained to validate a sample of the death certificates and/or to resolve questions.

To illustrate the challenge of classifying cause of death, consider the following case example from Kircher and Anderson:

A 65-year-old woman was first seen by her physician five years before her death when she complained of dyspnea and substernal chest pain precipitated by exertion. The electrocardiogram on a standardized exercise test demonstrated depression in the ST segments of 1.5 mV. The patient's symptoms were alleviated by a hydrochlorothiazide-trimterene combination (Dyazide) and sublingual nitroglycerine until nine months before death, when the frequency and severity of angina increased. Propranolol hydrochloride was prescribed.

One month before death and ten days after the onset of a flulike illness, the patient developed chills, fever, and pleuritic pain. An x-ray film of the chest revealed patchy consolidation of both lungs. The leukocyte count was 20 x 109/L (20,000/mm^3). Blood cultures were positive for pneumococci. Seventy-two hours after penicillin G potassium therapy was initiated, the symptoms subsided.

One month after the episode of pneumonia, the patient sustained a myocardial infarction. Five days after the patient's admission to the hospital, death occurred suddenly. An autopsy revealed severe coronary atherosclerosis, left anterior descending coronary artery thrombosis, acute myocardial infarction, left ventricular myocardial rupture, hemopericardium, and cardiac tamponade.

In this case, the immediate cause of death was rupture of the myocardium. The rupture was due to an acute myocardial infarction occurring five days before death. The underlying cause of death – the condition setting off the chain of events leading to the death – was chronic ischemic heart disease. The deceased had had this condition for at least five years before her death. Influenza and pneumonococcal pneumonia should also be shown as other significant conditions that contributed to death.

Instructions for coding cause of death on death certificates can be found on the web page for the National Vital Statistics System of the National Center for Health Statistics, CDC (http://www.cdc.gov/nchs/about/major/dvs/handbk.htm). As of August 2000, the web page included links for a tutorial by the National Association of Medical Examiners and various handbooks.

Psychiatric disorders – a special challenge

The challenges of classification of physical disease are formidable, but psychiatric disorders present an even greater challenge due to the difficulty of finding satisfactory answers to the most basic of questions, "what is a case?" (John Cassel, Psychiatric epidemiology. In: S. Arieti (ed), *American handbook of psychiatry*. 2nd ed. NY, Basic Books, 1974, vol. 2, 401-410; Kendell, *Arch Gen Psychiatry* 1988; 45:374-376). Despite a (modest) increase in resources and effort aimed at unraveling the etiology of these disorders, causal relationships have been very difficult to demonstrate. A key reason for the lack of progress may be problems with the definition of mental disorders (Jacob KS. The quest for the etiology of mental disorders. *J Clin Epidemiol* 1994;47:97-99).

Laboratory and other objectively measurable physiological signs have been a tremendous asset for defining and classifying diseases. Accordingly the need to rely almost exclusively on symptoms, behavioral observation, response to treatment, course, and outcome – manifestations that are more difficult to measure with reliability and precision – has put psychiatric nosology at a great disadvantage compared with physical illness. Although recent progress in psychiatric nosology, reflected in DSM III, III-R, and IV is believed to have improved reliability of diagnoses, the resulting diagnostic classifications are probably heterogeneous with respect to etiology. Subclassification based on biological and molecular variables, to take advantage of the significant advances in biology and biotechnology, and on refined measures of environmental and psychological variables may reveal etiologic associations that are masked by the current reliance on syndrome-based definitions (Jacob). On the other hand, if a disorder represents a "final common pathway", as has been argued with respect to unipolar major depressive disorder, then diverse etiologies could conceivably result in a biologically cohesive phenomenon.

Measuring accuracy in classification and detection

In general, any deviation between the (often-unknown) truly relevant biological entity and the result of the system used to define and detect or quantify it can be regarded as measurement error. Later in the course we will take up the concept of information bias, which deals with the effects of measurement error on study findings. Here, though, we will present the basic measures used in epidemiology to quantify accuracy of detection and classification methods. These measures can be applied to the detection of any entity, of course, whether it is a disorder, an exposure, or any characteristic. Besides their use in epidemiology in general, these measures are important for the selection and interpretation of diagnostic tests used in clinical practice.

If a condition or characteristic can be present or absent, then the accuracy of our system of detection and labeling can be assessed by its ability to detect the condition in those who have it as well as by its ability to correctly classify people in whom the condition is absent. Note that for a rare condition, overall accuracy [(a+d)/n in the table below] primarily reflects the correct identification of noncases, thus giving little information about the correct identification of cases. Also, overall accuracy ignores the fact that different kinds of errors have different implications.

Epidemiologists therefore employ separate, complementary measures for the correct classification of cases and of noncases. The basic measures are:

- ***Sensitivity*** – the proportion of persons who have the condition who are correctly identified as **cases**.
- ***Specificity*** – the proportion of people who do not have the condition who are correctly classified as **noncases**.

The definitions of these two measures of validity are illustrated in the following table.

Classification contingency table

		True status +	True status –		
Classified status	+	a	b	(a + b)	(Positive tests)
	–	c	d	(c + d)	(Negative tests)
Total		a + c (Cases)	b + d (Noncases)		

In this table:

Sensitivity (accuracy in classification of ***cases***) = a / (a + c)

Specificity (accuracy in classification of ***noncases*****)** = d / (b + d)

Sometimes the following terms are used to refer to the four cells of the above table:

a = True positive, TP – people with the disease who test positive

b = False positive, FP – people without the disease who test positive

c = False negative, FN – people with the disease who test negative

d = True negative, TN – people without the disease who test negative

However, these terms are somewhat ambiguous (note that "positive" and "negative" refer to the result of the test and not necessarily to the true condition). The relative costs (financial and human) of false negatives and false positives are key factors in choosing between sensitivity and specificity when a choice must be made. The more urgent is detection of the condition, the greater the need for sensitivity. Thus, a condition that has severe consequences if left untreated and which can be readily treated if detected early implies the need for a test with high sensitivity so that cases are not missed. A condition for which an expensive, invasive, and painful diagnostic workup will follow the results of a positive test implies the need for a test with high specificity, to avoid false positive tests.

Criterion of positivity and the receiver operating characteristic

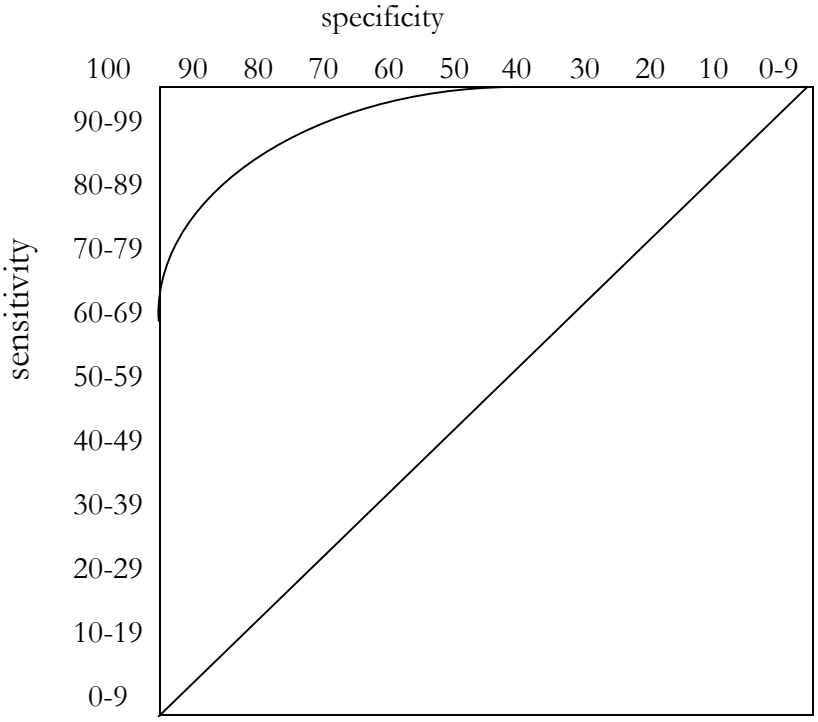
Often, however, the differences between cases and noncases are subtle in relation to the available classification system or detection methods. In such cases we can increase one type of accuracy only by trading it off against the other by where we set the "criterion of positivity", the cutpoint point used to classify test results as "normal" or "abnormal".

Screening for abnormal values of physiologic parameters is a typical situation. If we are attempting to classify people as diabetic or not based on their fasting blood glucose level, then we can set our cutpoint low in order to be sure of not missing diabetics (i.e., high sensitivity for detecting cases) but in doing so we will also include more people whose blood glucose falls at the upper part of the

distribution but are not diabetic (i.e., low specificity). If instead we choose a high cutpoint in order to avoid diagnosing diabetes when it is not present, then we are likely to miss diabetics (low sensitivity) whose blood glucose falls in the lower part of the distribution. Tests that involve rendering a judgment based on an image or specimen (e.g., mammography and cytology) involve similar, though less quantifiable, tradeoffs. As we shall see, in addition to the relative consequences of false negative and false positive tests, the decision of where to set the criterion of positivity should also take into account the prevalence of the condition in the population to be tested.

One useful technique for comparing the performance of alternative tests without first having to select a criterion for positivity and also for selecting a cutpoint is to graph the ***receiver/response operating characteristic (ROC)*** for each test (the concept and terminology come from engineering). The ROC shows the values of sensitivity and specificity associated with each possible cutpoint, so that its graph provides a complete picture of the performance of the test. For example, the sample ROC curve in the figure indicates that at 80% sensitivity, the test is about 95% specific. At 95% sensitivity, the specificity is only about 74%. If high sensitivity (e.g., 98%) is essential, the specificity will be only 60%.

Sample ROC curve



An ROC curve that consisted of a straight line from the lower left-hand corner to the upper right-hand corner would signify a test that was no better than chance. The closer the curve comes to the upper left-hand corner, the more accurate the test (higher sensitivity and higher specificity).

Predictive value

Sensitivity and specificity are, in principle, characteristics of the test itself. In practice, all sorts of factors can influence the degree of sensitivity and specificity that are achieved in a particular setting (e.g., calibration of the instruments, level of training of the reader, quality control, severity of the condition being detected, expectation of positivity). However, for any particular sensitivity and specificity, the yield of a test (accurate and inaccurate positive test results) will be determined by how widespread the condition is in the population being tested. The typical difficulty is that, since the number of people without the condition is usually much larger than the number with the condition, even a very good test can easily yield more false positives than true ones.

The concept of ***predictive value*** is used to assess the performance of a test in relation to a given frequency of the condition being sought. The ***positive predictive value*** (PPV) is defined as the proportion of people with the condition among all those who received a positive test result. Similarly, the ***negative predictive value*** is the proportion of people without the condition among all those who received a negative test result. Using the same table as before:

Classification contingency table

		True status +	True status –		
Classified status	+	a	b	(a + b)	(Positive tests)
	–	c	d	(c + d)	(Negative tests)
Total		a + c	b + d		
		(Cases)	(Noncases)		

Positive predictive value (PPV) = a / (a + b)

Negative predictive value (NPV) = d / (c + d)

Predictive value is an essential measure for assessing the effectiveness of a detection procedure. Also, since predictive value can be regarded as the probability that a given test result has correctly classified a patient, this concept is also fundamental for interpreting a clinical measurement or diagnostic test as well as the presence of signs or symptoms. The PPV provides an estimate of the probability that someone with a positive result in fact has the condition; the NPV provides an estimate that someone with a negative result does not in fact have the condition. (For a full discussion of the use of predictive value and related concepts in diagnostic interpretation, see a clinical epidemiology text, such as that by Sackett *et al.*)

In a clinical encounter prompted by symptoms, there is often a substantial probability that the patient has the condition, so both sensitivity and specificity are important in determining the proportion of cases and noncases among those who receive positive tests. However, in a screening

program in the general population, the specificity will typically dominate. Even with perfect sensitivity, the number of true cases cannot exceed the populatiion size multiplied by the prevalence, which is usually small. The number of false positives equals the false positive rate (1–specificity) multiplied by the number of nocases, which for a rare disease is almost the same as the population size. So unless the prevalence is greater than the false positive rate, the majority of test positives will not have the disease. For example, if only 1% of the population has the condition, then even if the specificity is 95% (false positive rate of 5%) the group who receive positive tests will consist primarily of noncases:

Cases detected (assume 100% sensitivity):

100% sensitivity x 1% with the condition = 1% of population

False positives:

95% specificity x 99% without the condition = 94.05% of population correctly classified, leaving 5.95% incorrectly labeled positive

Total positives:

1% + 5.95% = 6.95% of population

Proportion of positives who are cases (PPV) = 1% / 6.95% = 14%

In a population of 10,000 people, the above numbers become 100 (1%) cases, all of whom are detected, and 9,900 noncases, 595 of whom receive positive tests, for a total of 695 people receiving positive tests, 100 of whom have the condition. We will take up some of the these issues at the end of our discussion of natural history of disease.

The dependence of PPV on sensitivity, specificity, and prevalence can be expressed algebraically, as follows:

$$\text{PPV} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} = \frac{1}{1 + \dfrac{\text{False positives}}{\text{True positives}}}$$

$$\text{PPV} = \frac{1}{1 + \dfrac{(1 - \text{specificity})\,(1 - \text{prevalence})}{\text{sensitivity} \times \text{prevalence}}}$$

This expression shows that PPV is related to the ratio of false positives to true positives. The larger the ratio, the lower the PPV. If the condition is rare, then (1 – prevalence) is close to 1.0, and even with perfect sensitivity (sensitivity = 1.0), the ratio of false positives to true positives is no less than the ratio of (1 – specificity) [the false positive rate] divided by the prevalence. So for small

prevalences, even a small false positive rate (e.g., 1%) can reduce PPV substantially. Conversely, applying the test in a high prevalence population (e.g., prevalence 10%) can yield an acceptable PPV in spite of a much higher false positive rate (e.g., 10%). When a test is used for diagnostic purposes, the patient is suspected of having the condition, so the PPV for a positive result is much greater than when the same test is used for population screening.

Natural history of disease

Diseases and other phenomena of interest in epidemiology are processes. For example, the process by which bronchogenic lung cancer arises involves a progression over many years of the development of abnormal cells in the bronchial epithelium. Several grades of abnormality (metaplasia, mild dysplasia, moderate dysplasia, severe displasia) have been described. For the most part these abnormalities have the potential to disappear spontaneously or regress. However, in one or a number of cells the abnormality progresses to carcinoma *in situ* and then to invasive carcinoma. Depending upon the focus of investigation, the process can extend to very early stages. If our focus is on primary prevention, we might consider the process of smoking onset, usually in adolescence, the development of nicotine addiction and the smoking habit, and repeated attempts to quit. We might also consider the effects of tobacco marketing on smoking onset and maintenance, and the effects of legislation, litigation, competition, and investment opportunities on tobacco industry practices.

Thus, defining, observing, and measuring health and disease requires an appreciation of the concept of natural history – the evolution of a pathophysiologic process. "Natural" refers to the process in the absence of intervention. Natural history encompasses the entire sequence of events and developments from the occurrence of the first pathologic change (or even earlier) to the resolution of the disease or death. The natural history of a disease may be described through a ***staging classification***. Staging can aid in defining uniform study groups for research studies, determining treatment regimens, predicting prognosis, and in providing intermediate end-points for clinical trials.

Natural history therefore includes a ***presymptomatic*** period and a ***postmorbid*** period. Of particular interest for epidemiologists is the former, the period of time before clinical manifestations of the disease occur and, for infectious diseases, the period of time between infection and infectiousness. For non-infectious diseases, the term ***induction period*** refers to the "period of time from causal action until disease initiation" (Rothman and Greenland, p14). The induction period may be followed by a ***latent period*** (also called ***latency***), which is the "time interval between disease occurrence and detection" (Rothman and Greenland, p15). This distinction, though not made by all authors, is important for diseases that can be detected through screening tests, since the latent period represents the stage of the disease natural history when early detection is possible.

The distinction is also important for designing epidemiologic studies. Since the time of disease detection may be advanced through the application of screening and diagnostic tests, the number of cases detected can change with technology. Also, the collection of historical exposure data should be guided by a concept of when such exposure would have been biologically relevant. For a factor believed to contribute to the initiation of a disease, exposure must occur before that point. For a factor believed to contribute to promotion or progression of the condition, exposure can take place following initiation.

For infectious diseases, there are two markers of epidemiologic importance: disease detection and the onset of infectiousness. ***Incubation period*** refers to the "time from infection to development of symptomatic disease" (Halloran, p530). This term is sometimes applied to non-infectious diseases, but often without a precise meaning. The incubation period thus covers both the induction and latent periods as these are defined for non-infectious diseases. In contrast, the term latent period has a different meaning for infectious diseases, where it denotes "the time interval from infection to development of infectiousness" (Halloran, p530). Since an infected person may be infectious before developing symptoms, while symptomatic, or after disappearance of symptoms, there is no absolute relationship of incubation and latent periods for infectious disease. Relevant concepts are ***inapparent or silent infection*** (asymptomatic, either infectious or non-infectious) and ***carrier*** (post-symptomatic but still infectious) (Halloran, pp530-531).

Infectious disease	
Incubation	"time from infection to development of symptomatic disease" (Halloran, p530)
Latency	"the time interval from infection to development of infectiousness" (Halloran, p530)
Non-infectious disease	
Induction	"period of time from causal action until disease initiation" (Rothman and Greenland, p14)
Latency	"time interval between disease occurrence and detection" (Rothman and Greenland, p15)

Acute versus chronic diseases

Historically, disease natural histories have been classified into two broad categories: acute and chronic. ***Acute diseases*** (typically infections) have short natural histories. ***Chronic diseases*** (e.g., cancer, coronary heart disease, emphysema, diabetes) have long natural histories. So great has been the dichotomy of acute/infectious disease versus chronic/noninfectious disease that many epidemiologists and even departments of epidemiology are frequently regarded as one or the other.

In 1973 in the first Wade Hampton Frost Lecture, Abraham Lilienfeld regretted the concept of "Two Epidemiologies" and sought to emphasize the aspects in common between infectious and noninfectious epidemiology (see *Am J Epidemiol* 1973; 97:135-147). Others (e.g., Elizabeth Barrett-Connor, Infectious and chronic disease epidemiology: separate and unequal? *Am J Epidemiol* 1979;

109:245) have also criticized the dichotomy both in terms of its validity and its effect on epidemiologic investigation.

The growth of evidence for viral etiologies for various cancers (notably T-cell leukemias and cervical cancer) as well as other chronic diseases (e.g., juvenile onset diabetes mellitus and possibly multiple sclerosis) and for the central roles of immune system functions in chronic diseases demonstrates the importance of building bridges between the two epidemiologies. Also problematic for the identities acute = infectious and chronic = noninfectious are slow viruses, such as HIV. HIV may or may not produce a brief, flu-like syndrome within a week or so after infection. During the several weeks or months the host antibody response develops, and the virus enters a prolonged subclinical state during which the virus appears to remain quiescent. Many years may elapse until a decline in CD4 lymphocytes occurs and results in (chronic) immune deficiency.

Knowledge of the pathophysiology of early HIV infection is the basis for the Serologic Testing Algorithm for Recent HIV Seroconversion (STARHS, Janssen *et al*., 1998). The STARHS technique uses an assay whose sensitivity has been deliberately reduced. Specimens found to be HIV-positive in a sensitive assay are retested with the "de-tuned assay". Failure to detect antibody with the less sensitive assay most likely signifies that the infection was recently-acquired and the antibody response has not fully developed. Thus, the technique makes it possible to establish what proportion of HIV infections in a population occurred recently, indicating the level of continuing transmission.

Spectrum of disease

Diseases typically involve a spectrum of pathologic changes, some of which are considered disease states and some pre-disease states. The spectrum of disease concept has been studied, at the cellular and molecular level, for both coronary artery disease and cancer. Seeing more of the full spectrum or sequence can make us less certain at what point the "disease" has actually occurred.

Coronary artery disease:

Coronary artery disease pathogenesis is now understood in considerable detail (e.g., see Fuster *et al. N Engl J Med*, Jan 23, 1992;326(4):242 and Herman A. Tyroler, Coronary heart disease in the 21st century. *Epidemiology Reviews* 2000;22:7-13). "Spontaneous" atherosclerosis is initiated by chronic minimal (Type I) injury to the arterial endothelium, caused mainly by a disturbance in the pattern of blood flow in certain parts of the arterial tree. This chronic injury can also be potentiated by various factors, including hypercholesterolemia, infection, and tobacco smoke constituents.

Type I injury leads to accumulation of lipids and monocytes (macrophages). The release of toxic products by macrophages leads to Type II damage, which is characterized by adhesion of platelets. Growth factors released by macrophages, platelets, and the endothelium lead to the migration and proliferation of smooth-muscle cells, contributing to the formation of a "fibrointimal lesion" or a "lipid lesion". Disruption of a lipid lesion leads to Type III damage, with thrombus formation.

Small thrombi can contribute to the growth of the atherosclerotic plaque. Large thrombi can contribute to acute coronary syndromes such as unstable angina, myocardial infarction, and sudden ischemic death. Autopsy studies have revealed early, microscopic lesions in infants, though they regress. In adolescents, fatty streaks containing smooth-muscle cells with lipid droplets are observed. At this age fatty streaks are not surrounded by a fibrotic cap, which develops on some lesions in the 20's. Progression to clinically manifest, enlarging atherosclerotic plaques, such as those causing exertional angina, may be slow (probably in response to Type I and Type II injury) or rapid (in response to Type III injury). At what point in this process is "coronary artery disease" present?

Cancer:

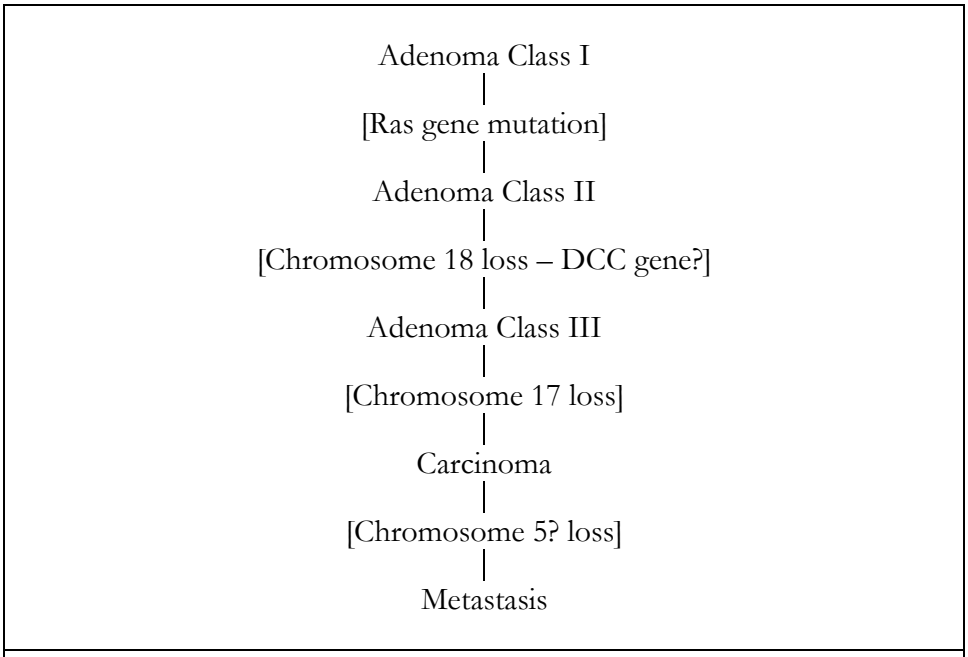
Cancer is also a multistage process, involving tumor initiation, promotion, conversion, and progression. Shields and Harris (Molecular epidemiology and the genetics of environmental cancer. *JAMA* August 7, 1991;266(5):681-687) describe the process as follows: "Tumor initiation involves the direct effects of carcinogenic agents on DNA, mutations, and altered gene expression. The attendant defects are involved in tumor promotion, whereby cells have selective reproductive and clonal expansion capabilities through altered growth, resistance to cytotoxicity, and dysregulation of terminal differentiation. Tumor promotion further involves an 'initiated' cellular clone that may also be affected by growth factors that control signal transduction. During this process, progressive phenotypic changes and genomic instability occur (aneuploidy, mutations, or gene amplification). These genetic changes enhance the probability of intiated cells transforming into a malignant neoplasm, the odds of which are increased during repeated rounds of cell replication. During tumor progression, angiogenesis allows for a tumor to grow beyond 1 or 2 mm in size. Ultimately, tumor cells can disseminate through vessels, invading distant tissues and establishing metastatic colonies." (681-682) When did "cancer" begin?

One of the cancers where understanding of natural history process has progressed to the identification of specific gene mutations is colon cancer. The process begins with initiation, e.g. chemical or radiation genetic damage to cell. It is now believed that alteration of a gene on chromosome 5 induces a transformation from normal colonic epithelium to a hyperproliferative epithelium. Initiated cells may then go through a series of distinct stages. The transformation process is enhanced by "promoters", which may be harmless in the absence of initiation. Stages that have so far been identified and the accompanying genetic alterations are shown in the accompanying figure. The progression from normal epithelium to cancer takes about ten years.

Stage of the cancer at diagnosis is influenced by various factors including screening and largely determines the outcome of therapy. Basic stages are: localized (in tissue of origin), regional spread (direct extension to adjacent tissues through tumor growth), and metastatic spread (tumor sheds cells that form new tumors in distant areas). Symptoms of various kinds develop according to the particular tissues and organs affected, and even the particular type of tumor cell (e.g., tumors in nonendocrine tissues can sometimes produce hormones).

Thus, the natural history of a disease can involve many, complex processes and developments long before the appearance of a clinical syndrome and even before the existence of the "disease" can be detected with the most sophisticated clinical tests. Moreover, particularly since some of the early

stages are spontaneously reversible, it is not always clear even theoretically when the "disease" itself is present.



(DCC = "deleted in colorectal cancer" gene) Correction of any one of chromosomes 5, 17, or 18 reduces malignancy, but only chromosome 18 restoration increases responsiveness to growth-inhibitory effects of transforming growth factor beta (TGF-beta).

Sources: Eric J. Stanbridge, Identifying tumor suppressor genes in human colorectal cancer. *Science* 5 Jan 1990; 247:12-13; Colorectal cancer: new evidence for the adenoma/carcinoma sequence. Editorial. *The Lancet* 25 July 1992; 340:210-211)

Understanding the natural history of diseases and other conditions of interest is fundamental for prevention and treatment, as well as for research. The effectiveness of programs for early detection and treatment of cancer, for example, depends upon the existence of an extended period where the cancer or a premalignant lesion is asymptomatic yet detectable and where treatment is more effective than after symptoms appear. In order to evaluate the efficacy of therapeutic interventions, knowledge of the natural history in the absence of treatment is crucial. These concepts will be illustrated by considering cancer screening procedures.

Natural history and screening

Population screening is defined as the application of a test to asymptomatic people to detect occult disease or a precursor state (*Screening in Chronic Disease*, Alan Morrison, 1985). The immediate

objective is to classify them as being likely or unlikely of having the disease under investigation. The goal is to reduce mortality and morbidity on the basis of evidence that earlier treatment improves patient outcomes. The design and evaluation of population screening programs depend crucially on the natural history of the disease in question.

For a screening program to be successful it must be directed at a suitable disease and employ a good test. Diseases for which screening may be appropriate are typically cancers of various sites (e.g., breast, cervix, colon, prostate), infectious diseases with long latency periods such as HIV and syphilis, and physiologic derangements or metabolic disorders such as hypertension, hypercholesterolemia, phenylketonuria, etc. What these conditions have in common is that they have serious consequences which can be alleviated if treatment is instituted early enough. The natural histories of these conditions involve a period of time when the condition or an important precursor condition (e.g., dysplasia) is present but during which there are no symptoms that will lead to detection.

Earlier in this topic we defined the latent period as the time between disease initiation and its detection. Cole and Morrison (1980) and Morrison (1985) refer to the total latent period as the ***total pre-clinical phase*** (TPCP). However, only a portion of the TPCP is relevant for screening – the period when the condition can be detected with the screening test. Cole and Morrison refer to this portion as the ***detectable pre-clinical phase*** (DPCP). The preclinical phases end when the patient seeks medical attention because of diagnostic symptoms. The DPCP is that part of the TPCP that begins when the screening test can detect the disease. Thus, the DPCP can be advanced if the screening test can be improved. The preclinical phase can be shortened by teaching people to observe and act promptly on early or subtle symptoms.

For a condition to be a suitable one for population screening, it must have a prolonged DPCP, thus providing ample time for advancing the date of disease detection and treatment. For a screening test to be suitable, it must be inexpensive, suitable for mass use, and without risk. It must have good sensitivity, so that the condition is not missed too often, which may give clients false reassurance. Moreover, the relevant sensitivity is for detecting the DPCP, rather than clinical disease, since it is the detection of the DCPC that provides the advantage from screening. The test must have excellent specificity, to avoid an excessive number of false positive tests. Importantly, the test must be able to maintain these attributes when administered and interpreted in volume in routine practice.

A major stumbling block in recommending population screening is the need to balance any benefit from early detection of cases against the expense, inconvenience, anxiety, and risk from the medical workups (e.g., colonoscopy, biopsy) that will be needed to follow-up positive tests on people who do not in fact have the condition. As demonstrated earlier, even a highly accurate test can produce more false positives than true ones when applied in a population where condition is very rare. Low positive predictive value (high proportion of false positives) has been a principal argument against HIV screening among applicants for marriage licenses, screening mammograms for women under age 50 years, and prostate cancer screening with prostate specific antigen (PSA).

(Although the test itself may be the same, it is important to distinguish between the use of a test for screening and its use for diagnosis. Since in the latter context the test has been motivated by the

presence of signs or symptoms and history, the prevalence of the condition among the test recipients is much greater, so that a positive test has a much higher positive predictive value. The term case-finding is sometimes used to refer to the application of the test to asymptomatic patients in a primary care setting. Case-finding likely assures effective follow-up for people receiving a positive test, though possible issues related to economic and personal costs of false positives remain.)

Criteria for early detection of disease through screening

Criteria to be met before screening for a given disease:

1. Natural history of disease must be understood
2. Effective treatment is available
3. A test is available by which the disease can be recognized in its pre-clinical phase
4. The application of screening makes better use of limited resources than competing medical activities

Evaluation of screening programs

Early outcomes for evaluating a screening program are stage of the disease and case fatality. If the screening is effective, the stage distribution for cases should be shifted towards earlier stages and a greater proportion of patients should survive for any given time period. Late outcomes are reduced morbidity and mortality. However, these outcome measures can all be affected by features of disease definition and natural history. Three potential pitfalls are lead time, length bias, and overdiagnosis.

Lead time is the amount of time by which screening advances the detection of the disease (i.e. the time between detection by a screening test and detection without a screening test). Even if the interval between the (unknown) biologic onset of the disease and death is unchanged, earlier detection will lengthen the interval between diagnosis and death so that survival appears lengthened. Lead time bias results when a screening program creates the appearance of delaying morbidity and mortality but in reality does not alter the natural history.

Length bias results if tumors are heterogeneous in respect to their aggressiveness, with slower growing tumors having a more favorable prognosis (or at least longer time to death). Slower growing tumors are more likely to be detected by screening, since they are present and asymptomatic longer (i.e., they have a longer DPCP) than are rapidly growing, aggressive tumors. So tumors detected by screening will overrepresent slow growing, hence survivable, tumors than will cancers detected because of appearance of symptoms (the latter cases are called "interval cases" because they are detected during the interval between screens).

Overdiagnosis results from the detection, by the screening test, of nonmalignant lesions that are judged to be malignant or to have malignancy potential. Prior to the use of the screening test, such lesions would not be detected, so their true prognosis may be unknown. If persons

with these apparently very early lesions are counted as having the disease, yet such lesions would not in any event progress to clinically-significant tumors, the survival experience of cases detected by screening will appear better. Overdiagnosis is a particular concern in evaluating the efficacy of prostate cancer screening.

Randomized trials, in which mortality is compared between a group offered screening and a group not offered screening (the classic study of this type is the Health Insurance Plan [HIP] trial of breast cancer screening) provide protection against these biases. But because they must usually be very large and of long duration, such trials are often difficult and very costly. The National Cancer Institute is currently conducting a very large (74,000 men and 74,000 women) and lengthy randomized trial to evaluate the effectiveness of screening for prostate, lung, colorectal, and ovarian cancers.

Both natural history and screening considerations come into play in such questions as the interpretation of secular changes in incidence and mortality. According to the NCI SEER (Surveillance, Epidemiology and End Results) Program, newly diagnosed cases of breast cancer increased between 1950 and 1979 at an annual rate of 1%, and between 1980 and 1984 at an annual rate of 3% (Breast cancer incidence is on the rise – but why? *JNCI* June 20, 1990; 82(12):998-1000). There has also been a "dramatic" upsurge in *in situ* breast cancer diagnosed since 1983. Breast cancer mortality overall was stable in the 1970s and began to fluctuate in the mid-1980s. Are the observed changes due to increased use of mammography? In support of that interpretation is the fact that among white women age 50 and older, localized disease has increased (i.e., a shift in the stage distribution) during the 1980s. There has also been a rapid increase in sales and installation of new mammography units during the 1980s, and the number of mammograms has risen dramatically. Or, could the observed changes be due to changes in risk factors (e.g., oral contraceptives, alcohol consumption, diet)? The observation of a striking increase in estrogen-receptor positive cancers suggests some biological change has occurred.

Another cancer where issues of natural history and early detection are of great importance is cancer of the prostate. The substantial (e.g., around 30% in men age 50 years and older) prevalence of previously undetected prostate cancer found at autopsy has demonstrated that many more men die with prostate cancer than from prostate cancer. Although "indolent" prostate cancers have the pathological features of cancer, if their growth is so slow that they will never become clinically manifest, should they be considered as the same disease as cancers of clinical importance? In addition, the lengthy natural history of most prostate cancers raises the concerns of lead time bias, length bias, and overdiagnosis for any observational approach to evaluating the efficacy of screening for early prostate cancer. In addition, there are major questions about the effectiveness of both existing modes of treatment and existing modes of early detection. Prostate cancer incidence **doubled** from 90 per 100,000 in 1985 to 185 per 100,000 in 1992, undoubtedly as a result of the dissemination of prostatic-specific antigen (PSA) screening. Meanwhile, prostate cancer mortality has decreased, though more modestly. These trends are consistent with the claim that screening with PSA reduces mortality, though the issue remains controversial for a number of reasons.

Bibliography

Lilienfeld and Lilienfeld – *Foundations of epidemiology*, Chapters 3-7; MacMahon & Pugh – *Epidemiology principles and methods*, Chapters 4, 7-10; Mausner & Kramer – *Epidemiology: an introductory text*, Chapter 1-2, 6. Kelsey, Thompson & Evans – *Methods in observational epidemiology*, pp. 23-31 and 46-53.

Chorba, Terence L.; Ruth L. Berkelman, Susan K. Safford, Norma P. Gibbs, Harry F. Hull. Mandatory reporting of infectious diseases by clinicians. *MMWR* 1990 (June 20); 39(9):1-6.

Cole, Phillip; Alan S. Morrison. Basic issues in population screening for cancer. *JNCI* 1980; 64:1263-1272.

Dubos, Rene. *Man adapting.* New Haven, CT, Yale, 1965.

Feinstein, Alvan R. The Blame-X syndrome: problems and lessons in nosology, spectrum, and etiology. *J Clinical Epidemiol* 2001;54:433-439.

Goodman, Richard A.; Ruth L. Berkelman. Physicians, vital statistics, and disease reporting. Editorial. *JAMA* 1987; 258:379-380.

Halloran, M. Elizabeth. Concepts of infectious disease epidemiology. In: Rothman and Greenland, *Modern Epidemiology* 2ed, Philadelphia, Lippincott-Raven, 1998, ch 27.

Israel, Robert A.; Harry M. Rosenberg, Lester R. Curtin. Analytical potential for multiple cause-of-death data. *Am J Epidemiol* 1986; 124:161-179. See also: Comstock, George W.; Robert E. Markush. Further comments on problems in death certification. 180-181.

Jacob KS. The question for the etiology of mental disorders. *J Clin Epidemiol* 1994;47:97-99.

Janssen RS, Satten GA, Stramer SL, et al. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. *JAMA* 1998; 280:42-48.

Kircher, Tobias; Robert E. Anderson. Cause of death: proper completion of the death certificate. *JAMA* 1987;258:349-352.

Kircher T, Nelson J, Burdo H. The autopsy as a measure of accuracy of the death certificate. *N Engl J Med* 1985; 313:1263-9.

Lindahl, B.I.B; E. Glattre, R. Lahti, G. Magnusson, and J. Mosbech. The WHO principles for registering causes of death: suggestions for improvement. *J Clin Epidemiol* 1990; 43:467-474.

Mirowsky, John and Catherine E. Ross. Psychiatric diagnosis as reified measurement. *J Health and Social Behavior* 1989 (March): 30:11-25 plus comments by Gerald L. Klerman (26-32); Marvin Swartz, Bernard Carroll, and Dan Blazer (33-34); Dan L. Tweed and Linda K. George (35-37); and rejoinder by Mirowsky and Ross (38-40).

Morrison, Alan S. *Screening in chronic disease.* NY, Oxford, 1985. (His chapter on screening in Rothman and Greenland is an excellent succinct presentation of concepts related to screening programs and their evaluation.)

National Vital Statistics System, National Center for Health Statistics, CDC. Various handbooks on completing death certificates. http://www.cdc.gov/nchs/about/major/dvs/handbk.htm

Percy, Constance; Calum Muir. The international comparability of cancer mortality data. *Am J Epidemiol* 1989; 129:934-46.

Percy, Constance; Edward Stanek III, Lynn Gloeckler. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *Am J Public Health* 1981; 71:242-250.

Rothman, Kenneth J. Induction and latent periods. *Am J Epidemiol* 1981; 114:253-9.

Sorlie, Paul D., Ellen B. Gold. The effect of physician terminology preference on coronary heart disease mortality: an artifact uncovered by the 9th Revision ICD. *Am J Public Health* 1987; 77:148-152.

Stallones, Reuel A. The rise and fall of ischemic heart disease. Scientific American 1980; 243:53-59.

Stanbridge, Eric J. Identifying tumor suppressor genes in human colorectal cancer. *Science* 5 Jan 1990; 247:12-13)

Temple LKF, McLeod RS, Gallinger S, Wright JG. Defining disease in the genomics era. *Science* 3 August 2001;293:807-808. See also letters by Byrne GI and Wright JG, 7 Sept 2001;293:1765-1766.

World Health Organization. *Manual of the international statistical classification of diseases, injuries, and causes of death.* Based on recommendations of the Seventh Revision Conference, 1955. Vol I. World Health Organization, Geneva, 1957.