An editorial on this Classic Article in medical computing appears elsewhere in this issue.

A MATHEMATICAL APPROACH TO MEDICAL DIAGNOSIS: APPLICATION TO CONGENITAL HEART DISEASE

HOMER R. WARNER, M.D., Ph.D., ALAN F. TORONTO, M.D., L. GEORGE VEASEY, M.D., AND ROBERT STEPHENSON, Ph.D.

An equation of conditional probability is derived to express the logical process used by a clinician in making a diagnosis based on clinical data. Solutions of this equation take the form of a differential diagnosis. The probability that each disease represents the correct diagnosis in any particular patient can be calculated. Sufficient statistical data regarding the incidence of clinical signs, symptoms, and electrocardiographic findings in patients with congenital heart disease have been assembled to allow application of this approach to differential diagnosis in this field. This approach provides a means by which electronic computing equipment can be used to advantage in clinical medicine.

Diagnosis of disease on the basis of clinical data is considered by the medical profession to be a subtle art that can be mastered only after years of careful study and extensive personal experience. Although rapid advances are being made in the development of new and improved methods for acquiring objective information from a patient concerning an illness, similar progress has not been made in analyzing and improving the logical process by which a diagnosis is deduced from this information. The present study was undertaken to find an explicit mathematical expression for this logical process, with the hope that such an expression might improve the accuracy of diagnosis in certain fields of medicine, lead to a more scientific approach to the teaching of medical diagnosis, and provide a means, with the help of an electronic computer, for relieving the physician of the task of storing and recalling for practical use in diagnosis an ever-increasing mass of statistical data. The derivation of such an equation is herein presented and its usefulness illustrated in its application to the diagnosis of congenital heart disease on the basis of clinical data.

Theory

That the logical process involved in medical diagnosis could be expressed as a problem in conditional probability (1) was suggested by Ledley and Lusted (2). The problem consists of estimating the likelihood or probability of event y_1 occurring in the presence of another event, x. In this paper the event y_1 is one disease among a series of diseases $y_1, y_2, \ldots y_k$, assumed to be mutually exclusive, and the event x is a set of clinical findings $x_1, x_2, \ldots x_j$, which will here be called symptoms even though physical signs and electrocardiographic findings are included. The probability of y_1 is defined by

$$\text{Equation 1: } P_{y_1} = \frac{N_{y_1}}{N_{(y_1, y_2, \ldots y_k)}}$$

where N_{y_1} is the number of times disease y_1 would occur in a large random sample of $N_{(y_1, y_2, \ldots y_k)}$ patients with diseases $y_1, y_2, \ldots y_k$. (P_{y_1}) is simply the incidence of disease y_1 in this subpopulation consisting only of people having one of these diseases. The probability of symptoms x occurring in a patient with disease y_1 is given by

$$\text{Equation 2: } P_{x|y_1} = \frac{N_{xy_1}}{N_{y_1}}$$

where N_{xy_1} is the number of patients with disease y_1 also having symptoms x. Dividing the numerator and denominator of the right-hand term by the size of the population $N_{(y_1, y_2, \ldots y_k)}$ results in

$$\text{Equation 3: } P_{x|y_1} = \frac{P_{xy_1}}{P_{y_1}}.$$

By the same reasoning the probability of disease y_1 occurring in the presence of symptom complex x may be written as

$$\text{Equation 4: } P_{y_1|x} = \frac{P_{xy_1}}{P_x}$$

where P_x is the probability of symptoms x occurring in any patient with one of those diseases. If these diseases are considered mutually exclusive, it follows that

$$\text{Equation 5: } P_x = \sum_{\text{all } k} P_{y_k} P_{x|y_k}$$

Combining equations 3, 4, and 5 results in

$$\text{Equation 6: } P_{y_1|x} = \frac{P_{y_1} P_{x|y_1}}{\sum_{\text{all } k} P_{y_k} P_{x|y_k}}$$

which is an expression of Bayes' rule for the probability of causes. Now, in fact, any symptom complex (x) may be represented as a series of independent symptoms $x_1, x_2, \ldots x_j$. Thus, the condition-

SYMPTOMS TO BE EVALUATED BY THE PHYSICIAN

Table 1

Code*	Symptom†
BW	{ x_1 = age 1 mo to 1 yr
BW	x_2 = age 1 to 20 yr
BW	x_3 = age >20 yr }
BW	{ x_4 = cyanosis, mild
BW	x_5 = cyanosis, severe (with clubbing)
BW	x_6 = cyanosis, intermittent
BW	x_7 = cyanosis, differential }
BW	x_8 = squatting
BW	x_9 = dyspnea
BW	x_{10} = easy fatigue
BW	x_{11} = orthopnea
BW	x_{12} = chest pain
BW	x_{13} = repeated respiratory infections
BW	x_{14} = syncope
BW	x_{15} = systolic murmur loudest at apex
B	x_{16} = diastolic murmur loudest at apex
B	{ x_{17} = systolic murmur loudest in left 4th interspace
B	x_{18} = diastolic murmur loudest in left 4th interspace
BW	x_{19} = continuous murmur loudest in left 4th interspace }
B	{ x_{20} = systolic murmur with thrill loudest in left 2nd interspace
B	x_{21} = systolic murmur without thrill loudest in left 2nd interspace
BW	x_{22} = diastolic murmur loudest in left 2nd interspace
BW	x_{23} = continuous murmur loudest in left 2nd interspace }
BW	x_{24} = systolic murmur loudest in right 2nd interspace
BW	x_{25} = diastolic murmur loudest in right 2nd interspace
BW	{ x_{26} = systolic murmur heard best over posterior chest
BW	x_{27} = continuous murmur heard best over posterior chest
BW	x_{28} = accentuated 2nd heart sound in left 2nd interspace
BW	x_{29} = diminished 2nd heart sound in left 2nd interspace }
BW	x_{30} = right ventricular hyperactivity by palpation
BW	x_{31} = forceful apical thrust
BW	x_{32} = pulsatile liver
BW	x_{33} = absent or diminished femoral pulsation
BW	{ x_{34} = ECG axis more than 110°
BW	x_{35} = ECG axis less than 0° }
BW	{ x_{36} = R wave greater than 1.2 mV in lead V_1
BW	x_{37} = R′ or qR pattern in lead V_1 }
BW	x_{38} = R wave greater than 2.0 mV in lead V_6
BW	x_{39} = T wave in lead V_6 inverted (no digitalis)
W	{ x_{40} = early diastolic murmur loudest at apex
W	x_{41} = late diastolic murmur loudest at apex }
W	{ x_{42} = holosystolic murmur loudest in left 4th interspace
W	x_{43} = midsystolic murmur loudest in left 4th interspace }
W	{ x_{44} = holodiastolic murmur loudest in left 4th interspace
W	x_{45} = early diastolic murmur loudest in left 4th interspace }
W	{ x_{46} = midsystolic murmur with thrill loudest in 2nd left interspace
W	x_{47} = holosystolic murmur with thrill loudest in 2nd left interspace
W	x_{48} = midsystolic murmur without thrill loudest in 2nd left interspace
W	x_{49} = holosystolic murmur without thrill loudest in 2nd left interspace }
BW	x_{50} = murmur louder than gr 3/6

*B indicates that the symptom was used on the brown check-off sheet; W indicates that the symptom was used on the white check-off sheet.
†Symptoms within braces are mutually exclusive and must be handled as special cases (see text).

DISEASES INCLUDED IN DIFFERENTIAL DIAGNOSIS

Table 2

y_1 = normal
y_2 = atrial septal defect without pulmonary stenosis or pulmonary hypertension*
y_3 = atrial septal defect with pulmonary stenosis
y_4 = atrial septal defect with pulmonary hypertension*
y_5 = complete endocardial cushion defect (A-V commune)
y_6 = partial anomalous pulmonary venous connections (without atrial septal defect)
y_7 = total anomalous pulmonary venous connections (supradiaphragmatic)
y_8 = tricuspid atresia without transposition
y_9 = Ebstein's anomaly of tricuspid valve
y_{10} = ventricular septal defect with valvular pulmonary stenosis
y_{11} = ventricular septal defect with infundibular stenosis
y_{12} = pulmonary stenosis, valvular (with or without probe-patent foramen ovale)
y_{13} = pulmonary stenosis, infundibular (with or without probe-patent foramen ovale)
y_{14} = pulmonary atresia
y_{15} = pulmonary artery stenosis (peripheral)
y_{16} = pulmonary hypertension,* isolated
y_{17} = aortic-pulmonary window
y_{18} = patent ductus arteriosus without pulmonary hypertension*
y_{19} = pulmonary arteriovenous fistula
y_{20} = mitral stenosis
y_{21} = primary myocardial disease
y_{22} = anomalous origin of left coronary artery
y_{23} = aortic valvular stenosis
y_{24} = subaortic stenosis
y_{25} = coarctation of aorta
y_{26} = truncus arteriosus
y_{27} = transposed great vessels
y_{28} = corrected transposition
y_{29} = absent aortic arch
y_{30} = ventricular septal defect without pulmonary hypertension*
y_{31} = ventricular septal defect with pulmonary hypertension*
y_{32} = patent ductus arteriosus with pulmonary hypertension*
y_{33} = tricuspid atresia with transposition

*Pulmonary hypertension is defined as pulmonary artery pressure ≧ systemic arterial pressue.

al probability ($P_{x|y_1}$) of symptom complex x occurring in disease y_1 must be the product of the probabilities of the individual symptoms that make up the set occurring in disease y_1. This is expressed in

Equation 7:

$$P_{x|y_1} = P_{x_1|y_1} P_{x_2|y_1} \ldots P_{x_j y_1}.$$

In order to clarify the meaning of independence of individual symptoms let us consider the case of two symptoms, x_a and x_b. It might be argued that for x_a to be truly independent of x_b, the probability of x_a must not be influenced by the presence of x_b; that is

$$\text{Equation 8: } P_{x_a|x_b} = P_{x_a}.$$

However, this can be true only if x_b is uniformly distributed throughout the population. This means that $P_{x_b|y_1} = P_{x_b|y_2} = P_{x_b|y_k} = 1$ and that x_b is of no diagnostic value. For this reason Equation 8 must be an inequality. In spite of this, these symptoms for present purposes are truly independent of each other as long as this inequality is due only to the non-uniform distribution of x_b in diseases y_1, $y_2, \ldots y_k$ and not due to a direct causal relationship between x_a and x_b. In the selection of symptoms to be used in a particular field, care must be taken to adhere to this criterion as closely as possible.

With use of Equation 7, we may rewrite Equation 6 in an expanded form as

Equation 9:

$$P_{y_1(x_1, x_2, \ldots x_j)} = \frac{P_{y_1} P_{x_1|y_1} \ldots P_{x_j|y_1}}{\sum_{\text{all } k} P_{y_k} P_{x_1|y_k} P_{x_2|y_k} \ldots P_{x_j|y_k}}.$$

With this expression it is possible to calculate the probability ($P_{y_1|x_1, x_2, \ldots x_j}$) that each disease $y_1, y_2, \ldots y_k$ exists in the presence of symptoms $x_1, x_2, \ldots x_j$ from statistical information concerning the incidence of each disease P_{y_1} in the population under consideration and the incidence of each of the patients' symptoms in each of these diseases ($P_{x_1|y_k}$, $P_{x_2|y_k}$, etc.). These statistical data, required for the right-hand term of Equation 9, may be compiled and stored in a form (punched cards, punched paper tape, or magnetic tape) that will make them readily available for an electronic digital computer to extract the pertinent numbers (depending on the symptoms presented by the patient) for carrying out the calculation called for by the equation.

Because Equation 9 uses only probabilities involving the symptoms actually present in the patient under consideration, the absence of a particular symptom does not influence the diagnosis. Thus, in order to make use of the fact that the absence of a symptom may have a bearing on the probability of a given disease being present, Equation 9 is modified to give

Equation 10:

$$P_{y_1|(x_1, \bar{x}_8, \ldots x_j)} = \frac{P_{y_1} P_{x_1|y_1} (1 - P_{x_8|y_1}) \ldots P_{x_j|y_1}}{\sum_{\text{all } k} P_{y_k} P_{x_1|y_k} (1 - P_{x_8|y_k}) \ldots P_{x_j|y_k}}$$

where the bar above x_8 in the initial term indicates that symptom x_8 is not present in the patient under consideration. Because $P_{x_8|y_1}$ represents the probability of symptom x_8 occurring in disease y_1, its complement $(1 - P_{x_8|y_1})$ must represent the probability of symptom x_8 not occurring in a patient with disease y_1. Thus, the absence of a symptom is treated as a discrete event when Equation 10 is used, and the probability that a symptom is absent can be obtained directly from the probability figure for the presence of the symptom.

Application to Congenital Heart Disease

Because the accuracy of a diagnosis of congenital heart disease based on clinical symptoms can be checked by cardiac catheterization and/or findings at surgery, and because relatively objective clinical findings can easily be obtained, this field was chosen for a pilot study. The list of symptoms and diseases used in this study, with their corresponding symbols, is shown in Tables 1 and 2. Statistical information concerning the incidence of each of these symptoms in each of these diseases is presented in Table 3. The numbers in the first column represent the incidence of each disease (P_y) in the subpopulation made up of patients referred to this laboratory in whom congenital heart disease was suspected. The rest

SYMPTOM-DISEASE MATRIX

Table 3

Diseases	Incidence	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
y_1	0.100	01	49	50	01	00	01	00	01	01	10	03	05	05	03	05	01	70	02	07	00
y_2	.081	10	50	50	02	01	02	00	01	35	50	05	02	40	01	02	02	30	20	02	05
y_3	.005	30	60	10	20	10	20	00	01	60	70	05	02	10	10	02	02	05	05	02	57
y_4	.001	10	20	70	30	10	25	00	01	80	90	05	05	15	10	02	02	15	20	02	05
y_5	.027	20	50	30	15	05	10	00	01	40	50	05	05	30	05	60	15	90	40	02	10
y_6	.005	10	40	50	01	01	01	00	01	15	20	01	05	05	01	02	02	20	02	02	02
y_7	.001	20	70	10	65	10	05	00	01	70	80	05	05	20	05	02	02	10	15	10	05
y_8	.018	50	48	02	30	65	01	00	10	80	90	20	05	15	10	02	05	65	05	05	20
y_9	.001	10	45	45	22	44	01	00	22	80	80	10	30	15	22	05	25	95	25	05	05
y_{10}	.054	40	55	05	25	25	10	00	30	75	90	05	05	10	20	02	02	20	02	05	65
y_{11}	.063	40	55	05	30	30	10	00	40	75	90	05	05	10	25	02	02	20	02	05	65
y_{12}	.045	20	70	10	01	01	01	00	01	50	65	01	01	01	10	02	02	10	02	05	70
y_{13}	.013	20	70	10	01	01	01	00	01	50	65	01	01	01	10	02	02	10	02	02	70
y_{14}	.014	90	09	01	10	90	00	00	80	90	99	05	10	05	35	02	02	40	05	05	01
y_{15}	.001	05	45	50	01	01	01	00	01	01	01	01	01	01	01	04	01	02	01	01	02
y_{16}	.013	10	45	45	01	01	01	00	01	70	95	40	10	10	10	01	01	30	05	01	01
y_{17}	.001	30	60	10	05	01	01	00	01	10	10	05	01	10	01	05	10	20	05	60	01
y_{18}	.072	20	40	40	01	01	01	00	01	20	20	10	01	10	05	05	15	10	02	50	02
y_{19}	.002	20	30	50	45	45	01	00	01	10	20	05	01	01	10	05	02	10	02	20	02
y_{20}	.008	20	50	30	01	01	01	00	01	50	50	40	05	10	10	80	20	10	10	02	05
y_{21}	.013	70	29	01	01	01	01	00	01	40	50	20	01	05	05	15	02	05	02	02	02
y_{22}	.001	70	29	01	01	01	01	00	01	30	30	30	80	15	20	05	01	01	01	01	01
y_{23}	.036	10	80	10	01	01	01	00	01	20	30	20	15	01	35	20	02	20	10	02	05
y_{24}	.009	10	80	10	01	01	01	00	01	20	30	20	15	01	35	20	02	20	10	02	05
y_{25}	.054	10	70	20	01	01	01	00	01	20	30	20	01	01	05	05	01	20	10	02	02
y_{26}	.005	50	40	10	30	60	01	00	15	15	30	05	01	20	10	02	02	70	02	02	10
y_{27}	.063	90	10	00	20	60	05	10	05	60	70	20	01	05	10	05	02	50	02	02	03
y_{28}	.001	30	30	30	30	05	10	00	01	10	20	01	01	01	01	05	02	70	02	02	06
y_{29}	.001	60	39	01	01	01	01	80	30	10	50	05	20	01	20	05	02	50	02	02	10
y_{30}	.252	15	70	15	01	01	01	00	01	20	30	05	01	15	05	05	20	95	05	02	10
y_{31}	.081	30	60	10	30	50	10	00	05	60	70	20	10	20	10	05	01	50	10	02	05
y_{32}	.005	30	40	30	01	01	05	50	01	20	30	10	01	10	05	02	02	10	10	02	02
y_{33}	.069	40	55	05	50	20	10	00	01	80	90	20	01	30	05	05	10	70	05	02	10

of Table 3 is a matrix with symptoms along the horizontal axis and diseases listed vertically. For instance, the number 0.02 at the intercept of symptom x_4 and disease y_2 represents $P_{x_4|y_2}$, the probability or incidence of mild cyanosis occurring in a patient with atrial septal defect without pulmonary hypertension. (In this study pulmonary hypertension is arbitrarily defined as pulmonary artery pressure equal to or greater than aortic pressure.)

Several things about this symptom-disease matrix require explanation. Listed among the diseases is a category called normal. The incidence of normal (P_{y_1}) in this study is 0.10, since 10% of the patients referred to this laboratory for heart catheterization are normal by physiologic studies, which include dye-dilution curves. The figures in the incidence column and symptoms x_1, x_2, and x_a (age) may vary from one population to the next, while the other data, which express the probability of each symptom in each disease, should remain constant from one laboratory to the next. Each of the probabilities in the matrix was determined by us from a careful review of published data of others, particularly Keith and coworkers,[3] review of data obtained from 1035 patients referred to this laboratory for diagnostic catheterization, and estimates based on the pathologic physiology of the defect in the case of rare defects in which adequate statistics were not available.

Notice that each patient is classified according to age into one of three categories—1 month to 1 year, 1 year to 20 years, and over 20 years of age. The patient's age is treated as a symptom. For instance, the number 0.70 occurring at the intercept of x_2 and y_{13} indicates that this symptom (age, 1 to 20 years) will occur in 70 of 100 patients with

ptoms

x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	x_{30}	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}	x_{39}	x_{40}	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}	x_{48}	x_{49}	x_{50}
01	00	01	01	15	05	10	03	01	01	01	02	02	02	02	02	01	00	02	70	04	03	00	00	80	05	10
01	01	01	01	60	01	80	01	01	01	70	05	05	85	02	02	01	02	01	30	02	20	05	01	90	01	60
01	01	01	02	30	15	40	01	05	01	85	05	20	70	02	02	01	01	01	05	01	05	60	01	38	01	70
01	01	01	01	95	01	50	01	65	01	85	05	20	70	02	02	01	02	01	15	20	02	05	01	40	01	40
01	01	01	01	70	02	40	10	10	01	05	70	05	85	02	02	15	01	85	05	02	20	02	20	20	20	80
01	01	10	15	40	02	10	01	01	01	15	02	02	15	02	02	02	02	02	20	02	02	02	02	60	02	30
01	01	10	15	85	02	80	01	01	01	90	02	25	75	02	02	02	02	30	10	01	30	05	01	80	02	70
01	01	01	01	02	60	01	20	30	01	02	90	02	02	90	10	05	02	50	15	05	02	20	20	20	20	50
01	01	01	01	02	35	10	20	10	01	10	02	02	60	02	02	25	25	45	45	25	25	15	15	05	05	50
02	02	10	15	10	00	20	01	02	01	95	02	85	10	02	02	02	02	20	05	02	02	60	05	25	05	90
02	02	10	15	10	60	20	01	02	01	95	02	85	10	02	02	02	02	20	05	02	02	60	05	25	05	90
02	02	01	01	10	60	20	01	05	01	95	02	85	10	02	02	01	01	01	10	02	02	68	01	25	01	80
02	02	01	01	10	60	20	01	05	01	95	02	85	10	02	02	01	01	01	10	01	01	68	01	25	01	80
02	02	10	10	01	90	20	01	02	01	95	02	85	10	02	02	02	01	30	40	02	05	01	01	02	02	20
20	02	50	05	10	02	10	01	01	01	10	02	10	02	02	02	01	01	02	02	01	00	02	01	25	02	60
02	02	02	02	95	00	30	01	10	01	95	02	90	05	02	02	01	01	01	30	15	05	02	02	05	02	20
02	02	02	02	70	01	20	40	01	01	01	15	02	02	60	05	10	02	10	20	05	02	02	02	10	05	75
02	02	03	05	50	01	20	40	02	01	02	10	02	02	50	05	10	02	05	10	02	02	05	02	20	10	85
01	01	05	70	05	05	20	01	01	01	05	05	02	02	02	02	02	02	10	10	02	02	02	02	10	10	30
02	02	01	01	50	01	20	05	02	01	50	02	10	40	02	02	20	20	10	10	10	10	05	05	10	10	70
10	02	01	01	20	02	10	50	02	01	05	10	05	05	40	90	02	02	10	10	02	02	02	02	05	05	10
01	01	01	01	20	02	01	05	01	01	05	10	05	05	20	90	01	01	01	01	01	01	01	01	01	01	10
95	05	01	01	20	10	01	40	01	05	05	15	02	02	70	15	02	02	02	20	10	02	05	01	05	01	90
95	05	01	01	20	10	01	40	01	05	05	15	02	02	70	15	02	02	02	20	10	02	05	01	05	01	90
15	10	80	15	10	10	01	30	01	99	05	05	02	02	40	04	01	01	05	20	10	02	02	02	10	05	65
02	02	05	10	40	10	30	05	01	01	30	10	40	10	20	05	02	02	40	40	02	02	10	10	10	10	40
05	02	01	01	20	10	20	20	02	02	40	20	30	05	20	05	02	02	30	30	02	02	03	03	10	10	50
05	02	01	01	20	10	10	10	01	01	20	10	10	10	10	10	02	02	30	30	02	02	05	05	30	30	60
05	02	01	01	90	02	40	05	01	10	70	05	80	05	10	05	02	02	30	30	02	02	10	10	30	30	20
02	05	01	01	30	02	05	30	01	01	30	10	05	05	15	05	20	02	92	05	05	01	01	10	01	10	85
02	05	01	01	90	02	30	05	05	01	70	05	75	15	10	05	01	01	30	30	10	02	01	05	01	05	50
02	02	02	02	90	02	30	05	05	01	70	05	75	15	10	05	02	02	10	10	02	02	02	02	20	20	20
02	01	01	01	30	10	01	20	30	01	02	90	02	02	90	10	10	02	30	30	05	05	10	10	30	30	50

TERMS TO BE USED IN EQUATION 10 IN CASES OF MUTUALLY EXCLUSIVE SYMPTOMS

Table 4

Check sheet	Symptoms	Symptom present	Term to Be Used
B/W	x_1-x_3	x_1	P_{x_1}
		x_2	P_{x_2}
		x_3	P_{x_3}
B/W	x_4-x_7	x_4	P_{x_4}
		x_5	P_{x_5}
		x_6	P_{x_6}
		x_7	P_{x_7}
		none	$(1-P_{x_4}-P_{x_5}-P_{x_6}-P_{x_7})$
B/W	$x_{26}-x_{27}$	x_{26}	$P_{x_{26}}$
		x_{27}	$P_{x_{27}}$
		neither	$(1-P_{x_{26}}-P_{x_{27}})$
B/W	$x_{28}-x_{29}$	x_{28}	$P_{x_{28}}$
		x_{29}	$P_{x_{29}}$
		neither	$(1-P_{x_{28}}-P_{x_{29}})$
B/W	$x_{34}-x_{35}$	x_{34}	$P_{x_{34}}$
		x_{35}	$P_{x_{35}}$
		neither	$(1-P_{x_{34}}-P_{x_{35}})$
B/W	$x_{36}-x_{37}$	x_{36}	$P_{x_{36}}$
		x_{37}	$P_{x_{37}}$
		neither	$(1-P_{x_{36}}-P_{x_{37}})$
B	$x_{17}-x_{19}$	x_{17}	$P_{x_{17}}(1-P_{x_{18}})$
		x_{18}	$P_{x_{18}}(1-P_{x_{17}})$
		x_{19}	$P_{x_{19}}$
		none	$(1-P_{x_{17}})(1-P_{x_{18}})(1-P_{x_{19}})$
		x_{17}, x_{18}	$P_{x_{17}} P_{x_{18}}$
B	$x_{20}-x_{23}$	x_{20}	$P_{x_{20}}(1-P_{x_{22}})$
		x_{21}	$P_{x_{21}}(1-P_{x_{22}})$
		x_{22}	$P_{x_{22}}(1-P_{x_{20}}-P_{x_{21}})$
		x_{23}	$P_{x_{23}}$
		x_{20}, x_{22}	$P_{x_{20}} P_{x_{22}}$
		x_{21}, x_{22}	$P_{x_{21}} P_{x_{22}}$
		none	$(1-P_{x_{20}}-P_{x_{21}})(1-P_{x_{22}})(1-P_{x_{23}})$
W	$x_{19}; x_{42}-x_{46}$	x_{19}	$P_{x_{19}}$
		x_{42}	$P_{x_{42}}(1-P_{x_{44}}-P_{x_{45}})$
		x_{43}	$P_{x_{43}}(1-P_{x_{44}}-P_{x_{45}})$
		x_{44}	$P_{x_{44}}(1-P_{x_{42}}-P_{x_{43}})$
		x_{45}	$P_{x_{45}}(1-P_{x_{42}}-P_{x_{43}})$
		x_{42}, x_{44}	$P_{x_{42}} P_{x_{44}}$
		x_{43}, x_{44}	$P_{x_{43}} P_{x_{44}}$
		x_{42}, x_{45}	$P_{x_{42}} P_{x_{45}}$
		x_{43}, x_{45}	$P_{x_{43}} P_{x_{45}}$
		none	$(1-P_{x_{19}})(1-P_{x_{42}}-P_{x_{46}})\ (1-P_{x_{44}}-P_{x_{45}})$
W	$x_{40}-x_{41}$	x_{40}	$P_{x_{40}}$
		x_{41}	$P_{x_{41}}$
		neither	$(1-P_{x_{40}}-P_{x_{41}})$
W	$x_{22}-x_{23}$ $x_{46}-x_{49}$	x_{22}	$P_{x_{22}}(1-P_{x_{46}}-P_{x_{47}}-P_{x_{48}}-P_{x_{49}})$
		x_{46}	$P_{x_{46}}(1-P_{x_{22}})$
		x_{47}	$P_{x_{47}}(1-P_{x_{22}})$
		x_{48}	$P_{x_{48}}(1-P_{x_{22}})$
		x_{49}	$P_{x_{49}}(1-P_{x_{22}})$
		x_{46}, x_{22}	$P_{x_{46}} P_{x_{22}}$
		x_{47}, x_{22}	$P_{x_{47}} P_{x_{22}}$
		x_{48}, x_{22}	$P_{x_{48}} P_{x_{22}}$
		x_{49}, x_{22}	$P_{x_{49}} P_{x_{22}}$
		x_{23}	$P_{x_{23}}$
		none	$(1-P_{x_{22}})(1-P_{x_{23}})$ $(1-P_{x_{46}}-P_{x_{47}}-P_{x_{48}}-P_{x_{49}})$

infundibular pulmonary stenosis who come to this laboratory. In this way, then, the fact is recognized that the age of the patient does influence the probability of a given diagnosis.

Since the patient can belong in only one of the three age groups, these three "symptoms" cannot be considered independent of one another. Thus, if the patient's age is between 1 and 20 years, P_{x_2} is used in Equation 10 but the complement of the probability for the other two age groups is not used in this case.

Furthermore, it is important that care be taken not to include in the list of symptoms any two symptoms that invariably occur together, since this strongly suggests interdependence and a causal relationship between them. For instance, clubbing of the fingers was not included as a separate symptom since it occurs in the same patients with congenital heart disease who have evidence of severe cyanosis. Instead, it is included as part of the definition of severe cyanosis. Inclusion of redundant (interdependent) symptoms would result in an unreal increase in the probability of those diseases having a high incidence of these symptoms when these symptoms are present, and a falsely low probability when these symptoms are absent.

There are other symptoms in the list that are mutually exclusive. For instance, the existence of x_5 excludes by definition x_4, x_6, and x_7. Thus, it would be an error to consider the absence of x_4, x_6, and x_7 as additional pieces of information once x_5 is known to be present. On the other hand, the absence of x_4 through x_7 in a particular case (no cyanosis) is an important fact and must be recognized by using in Equation 10 the complement of the sum of the probabilities of each of these symptoms occurring in the disease in question, which is $1 - P_{x_4|y_1} - P_{x_5|y_1} - P_{x_6|y_1} - P_{x_7|y_1}$.

Groups of mutually exclusive symptoms are indicated by braces in Table 1, and a complete list

TEST CASE ILLUSTRATING EFFECT OF INCLUDING NEGATIVE INFORMATION

Table 5

	Diagnosis with Equation 9		Diagnosis with Equation 10		Diagnosis with Equation 10 and without x_1	
Symptom	**Disease**	**Probability**	**Disease**	**Probability**	**Disease**	**Probability**
x_3	y_{11}	0.33	y_{12}	0.62	y_{12}	0.73
x_{10}	y_{10}	0.28	y_{13}	0.21	y_{13}	0.24
x_{11}	y_{16}	0.11	y_{10}	0.07	y_{10}	0.02
x_{29}	y_{12}	0.14	y_{11}	0.04		
x_{34}	y_{13}	0.04	y_{16}	0.03		
x_{36}						
x_{43}						
x_{48}						

of mutually exclusive symptoms, together with instructions about what data should be used in solving Equation 10 in any particular case, is given in Table 4.

Use of the Computer

Because of the large number of calculations required to make each diagnosis in the example (congenital heart disease) used in this paper, it is necessary to use a digital computer if Equation 10 is to be solved in a practical way. This equation can be solved by almost any general-purpose electronic digital computer that has the capability of "floating decimal point" operation. The incidence of each symptom in each disease shown in the matrix is transferred to punch cards. These disease cards, together with cards that contain the program telling the computer what operations to perform, are transferred into the computer memory by a card-reading machine. Another punched card is prepared from a check-off list of symptoms on which the physician, after examination of the patient, has marked the symptoms presented by the patient. (X-ray data are not presented in this paper but are being evaluated for inclusion in the symptom list at the present time.)

From this information, the computer then calculates, with use of Equation 9 or 10, the probability of each of the 33 congenital heart diseases being present in the patient under consideration. The diseases with probability greater than 1% are printed out at the end of the calculation, together with their respective probabilities. Two symptom lists are checked off by the clinician after examination of each patient. On one list (brown sheet) murmurs are described only as to timing and location, while on the other list (white sheet) the time course of intensity of the murmurs is included (Table 1). Equation 10 is solved with each of these sets of symptoms, and the resulting differential diagnoses are compared. Although the calculation based on the white sheet often gave a higher probability to the correct diagnosis, this was not consistently the case, particularly in instances in which classification of the time course of murmur intensity was difficult even with the help of a phonocardiogram. The point to be made here is that in applying this approach to diagnosis a compromise must be reached between two alternatives: the desirability of using as much information as possible, and the limitations in accuracy with which the more detailed information can be observed in the patient and the necessary statistical data can be obtained.

Example

The case shown in Table 5 illustrates the effect of using both positive and negative information in making a diagnosis. The list of symptoms indicates that the patient was over 20 years of age and complained of easy fatigue and orthopnea; his pulmonary second sound was diminished; his electrocardiogram exhibited an axis greater than 110° and an R wave greater than 1.2 mV in lead V_1; and by phonocardiogram he had a midsystolic murmur, without a thrill, which was of equal intensity in the pulmonary (second left interspace) and the precordial (fourth left interspace) area. Calculation of the probabilities for each disease with use only of the positive information (Equation 9) resulted in a higher probability for tetralogy of Fallot (y_{10} and y_{11}) than for isolated pulmonary stenosis (y_{12} and y_{13}). However, when both positive information and negative information were taken into account, as when Equation 10 is used, the probability of isolated pulmonary stenosis became 0.83, while the probability of tetralogy of Fallot was only 0.11. This patient was later found to have y_{12} both by physiologic studies and at surgery.

Also illustrated in Table 5 is the way in which this approach can be used to evaluate the contribution made toward a diagnosis by any given symptom. Here the calculation has been carried out with and without the symptom of orthopnea (x_{11}). Had this patient not complained of orthopnea the probability of isolated pulmonary stenosis (y_{12} and y_{13}) would have been 0.97, as compared with 0.83 when orthopnea was considered present. This, of course, results from the fact that orthopnea rarely occurs in patients with y_{12} or y_{13}. Since the presence or absence of just one symptom may make a real difference in the differential diagnosis, as in this instance, it is apparent that each symptom on the list must be accurately evaluated in every case if the correct probabilities are to be calculated. For this reason, only the most objective symptoms should be included in the definition of the original list for any study, even if this must be done at some sacrifice of detail.

In the case of the present study we are under the impression from our experience to date that symptoms 10 through 14 detract from the accuracy of diagnosis as often as they contribute, because of the difficulty involved in assessing their actual presence or absence in many cases, as well as the inaccuracy of the available statistical data regarding the incidence of these symptoms in each of these diseases. These five symptoms might well be eliminated from the list.

Evaluation of Experience to Date

Because the differential diagnosis obtained with this approach represents an estimation of probabilities in which the statistical data of Table 3 are used, it is impossible from a limited number of cases to evaluate its accuracy. However, it is apparent from our experience to date with 36 cases that the most probable diagnosis estimated with Equation 10 agrees with the actual diagnosis made by physiologic studies and observation at surgery at least as often as does the most probable diagnosis estimated by three experienced cardiologists from the same clinical information. Furthermore, the differential diagnosis resulting from solution of the equation is frequently more complete and, in retrospect, often appears more logical to the clinicians than the differential diagnosis listed by each of them before seeing the equation's prediction.

It must be emphasized that Equation 10 was derived directly from the definition of conditional probability. Thus, any evaluation of the accuracy of the predictions made by this approach should be considered as testing the adequacy of the matrix of statistical data and not of the equation. Given the correct original data matrix and accurate observations of the patient, the calculated probabilities will be correct. Final refinement of the present data matrix must await the accumulation of sufficient data for calculation of new probabilities ($P_{x_j|y_k}$ Since the presence or absence of each symptom is determined in each case, and follow-up information almost invariably yields the diagnosis with certainty, the data for satisfactory recalculation of symptom incidence are routinely accumulating. The computer will be used to recalculate its own data matrix when the amount of data is sufficient.

Aids to Teaching

That an explicit expression of the logic used in medical diagnosis has potential usefulness as a tool for teaching diagnosis to medical students and physicians seems apparent. The approach here presented provides a framework within which any diagnostic problem can be formulated and critically analyzed.

Often the very act of attempting to formulate the problem in terms required for application of Equation 10 results in new insight by providing answers to such questions as:

1. What is the exact definition of each symptom and each disease?

2. Are certain symptoms interdependent and others mutually exclusive?

3. What symptoms are important determinants of the diagnosis and what symptoms are unimportant?

A solution of Equation 10 for any given set of symptoms provides an objective, reproducible standard against which students can check the accuracy of their own deductions from these symptoms. How modifying the symptom set in any desired fashion affects the differential diagnosis can be readily observed.

This approach to the teaching of diagnosis of congenital heart disease is in current use at this hospital and has met with enthusiastic acceptance by medical students.

Appendix

To illustrate the use of Equation 10, consider the simple case of a population consisting of just two diseases (y_1 and y_2) and three independent symptoms (x_1, x_2, and x_3). The relative incidence of these two diseases and the probability of each symptom in each disease are shown in the matrix below.

	Incidence	x_1	x_2	x_3
y_1	0.23	0.1	0.7	0.6
y_2	0.77	0.8	0.2	0.5

If the patient to be diagnosed presents with symptoms x_1 and x_3, Equation 10 would be solved with use of the following numbers to make the diagnosis:

$$P_{y_1|(x_1, \bar{x}_2, x_3)} = \frac{0.23(0.1)(1-0.7)(0.6)}{0.23(0.1)(1-0.7)(0.6) + 0.77(0.8)(1-0.2)(0.5)} = 0.016$$

and

$$P_{y_2|(x_1, \bar{x}_2, x_3)} = \frac{0.77(0.8)(1-.2)(0.5)}{0.23(0.1)(1-0.7)(0.6) + 0.77(0.8)(1-0.2)(0.5)} = 0.984$$

[Adapted from JAMA (1961; 177(3): 177–183) with the permission of the publisher.]

References

1. Feller, W., *An Introduction to Probability Theory and Its Application*, New York: John Wiley & Sons, Inc., 1960.
2. Ledley, R.S., and Lusted, L.B., Use of Electronic Computers to Aid in Medical Diagnosis, *Proc Inst Radio Engineers* 47: 1970–1977 (Nov.) 1959.
3. Keith, J.D., and others: *Heart Disease in Infancy and Childhood*, New York: The Macmillan Co., 1958.